Title
Statistical Analysis of Infectious Diseases in Nursing and Genomic Data

Author
Toyama, Joy

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Statistical Analysis of Infectious Diseases in Nursing and Genomic Data

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Public Health

by

Joy Toyama

2019

ABSTRACT OF THE DISSERTATION

Statistical Analysis of Infectious Diseases in Nursing and Genomic Data

by

Joy Toyama
Doctor of Public Health
University of California, Los Angeles, 2019
Professor Christina Michelle Ramirez, Chair

In a variety of settings, including the medical field, it is common for the number of variables gathered to far exceed the sample size. Along with a high dimension, many of these included variables are often correlated. This can pose problems for traditional methods. Much of the time, the data cannot be utilized completely as is, but instead requires previous research to guide researchers to choose relevant predictors prior to model selection. Traditional methods such as logistic regression and mixed models cannot necessarily converge and struggle with identifiability when the number of measurements collected approach or become larger than the number of patients in the study. Machine-learning techniques, including Random Forests and the newly developed Fuzzy Forests method, can accommodate data with high dimensionality. We concentrate on decision trees in particular because of their relative ease of use, availability and predictive ability. Random Forest is a widely used, parallelizable and computationally efficient method; however it does not acknowledge any correlation between variables leading to a preference for correlated predictors. Fuzzy Forest, on the other hand, explicitly explores the correlation structure among the variables, leading to unbiased variable importance measures. Fuzzy Forest, along with Random Forest, is utilized in three applications; smoking cessation in health care workers, re-arrest among homeless ex-offenders and genetic predictors of lithium response in individuals with Bipolar disorder.

The dissertation of Joy Toyama is approved.

Karabi Nandy

David Elashoff

Thomas R. Belin

Christina Michelle Ramirez, Committee Chair

University of California, Los Angeles

2019

To my parents and my sister, for their unwavering faith in me and constant support over the years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deep gratitude to Professor Christina Ramirez for her support, insight, and endless help in the preparation of this dissertation. I would also like to thank Professor Nandy, Professor Elashoff, and Professor Belin for kindly serving on my doctoral committee and providing suggestions and support during the development of this project.

VITA

2004	B.S. (Mathematics and Statistics), University of Washington, Seattle, WA.
2004-2006	Graduate Student Teaching Assistant, Department of Statistics, Oregon State University, Corvallis, OR.
2006	M.S. (Statistics), Oregon State University, Corvallis, OR.
2006-2008,2013	Special Reader, Department of Biostatistics, UCLA, Los Angeles, CA.
2008	M.S. (Biostatistics), University of California Los Angeles, Los Angeles, CA.
2008–Present	Graduate Student Researcher, School of Nursing, UCLA, Los Angeles, CA.
2015–2016	Special Reader, School of Nursing, UCLA, Los Angeles, CA.

PUBLICATIONS

A Nyamathi, B Salem, E Hall, T Oleskowicz, M Ekstrand, K Yadav, **J Toyama**, S Turner, Susan M Faucette(2017). Violent crime in the lives of homeless female ex-offenders. *Issues in mental health nursing*, 38(2), 122-131.

A Esguerra-Gonzales, M Ilagan-Honorio, S Fraschilla, P Kehoe, AJ Lee, T Marcarian, K Mayol-Ngo, PS Miller, J Onga, B Rodman, D Ross, S Sommer, S Takayanagi, **J Toyama**, F Villamor, SS Weigt, A Gawlinski(2013). CNE article: pain after lung transplant: high-frequency chest wall oscillation vs chest physiotherapy. *Am J Crit Care*22(2):115-124.

A Esguerra-Gonzales, M Ilagan-Honorio, P Kehoe, S Fraschilla, AJ Lee, A Madsen, T Marcarian, K Mayol-Ngo, PS Miller, J Onga, B Rodman, D Ross, Z Shameem, K Nandy , **J Toyama**, S Sommer, C Tamonang, F Villamor, SS Weigt, A Gawlinski(2014). Effect of high-frequency chest wall oscillation versus chest physiotherapy on lung function after lung transplant. *Appl Nurs Res.*27(1):59-66.

K De Azambuja, P Barman, **J Toyama**, D Elashoff, GW Lawson, LK Williams, K Chua, D Lee, JJ Kehoe, A Brodkorb, R Schweibert, S Kitchen, A Bhimani, DJ Wiley(2014). Validation of an HPV16-mediated carcinogenesis mouse model. *In Vivo*28(5):761-767.

NA Pike, CA Okuhara,**J Toyama**, BP Gross, WJ Wells, VA Starnes(2015). Reduced pleural drainage, length of stay, and readmissions using a modified Fontan management protocol. *J Thorac Cardiovasc Surg*150(3):481487.

SV Godbole, K Nandy, M Gauniyal, P Nalawade, S Sane, S Koyande, **J Toyama**, A Hedge, P Virgo, K Bhatia, RS Paranjape, AR Risbud, SM Mbulaiteye, RT Mitsuyasu(2016). HIV and cancer registry linkage identifies a substantial burden of cancers in persons with HIV in India. *Medicine*95(37).

CHAPTER 1

Introduction

Nursing bridges the gap between doctors and patients, not only in terms of care but also of information. Nurses not only help during surgeries and implementing new hospital wide policy changes but are critical to elevating the quality of life of patients. With such a diversity of settings and opportunities to create changes in the medical field, nurses have the potential to create and guide their corresponding specialized fields forward. The wealth of data generated through electronic medical records and through patient responses on questionnaires offer great promise of individualize health care and improved] patient outcomes. This deluge of information often is of high dimension and also is often correlated. This can pose problems with traditional methods. Much of the time, the data cannot be utilized completely as is, but instead requires previous research to guide researchers to choose relevant predictors prior to model selection. Traditional methods such as logistic regression and mixed models cannot necessarily converge and struggle with identifiability when the number of measurements collected approach or become larger than the number of patients in the study. Unfortunately this commonly occurs among nursing data. The following explores two situations where a number of measurements were taken and that traditional methods were used, but where alternative methods may have been able to offer additional guidance.

Machine-learning techniques along with traditional techniques can help shed light in these situations. We concentrate on decision trees in particular because of their relative ease of use, availability and predictive ability. Random Forests also have favorable predictive profiles when compared to other methods such as support vector machines and neural networks

([70], [49]). With the usage of R and SAS that have packages or add on packages available to the user, random forest is widely used, parallelizable and computationally efficient.

The next sections will present selected decision-tree based machine-learning methods, CART, Random Forest, Conditional inference trees, and Fuzzy Forests. Main concepts from each method will be presented from each of these ensemble methods.

1.1 CART

The base learner of our decision tree forest is Classification and Regression Trees (CART). The applications and extensions presented in subsequent chapters will be limited to the situation where the outcome is categorical, although it extends easily to the regression setting with continuous outcomes. When the outcome is categorical and made up of a set of classes, a classification tree can be used to produce a tree or model to predict what outcome class results from each of the branches or partitions in the data. In classification tree based methods, the purpose of a tree is to provide a sequence of decision rules that can be used to partition the observations such that each partition makes observations within the contained area as similar as possible.

1.1.1 Characteristics of a Tree

Each tree is comprised of nodes and branches where each node consists of a subset of the observations and a branch further partitions the data. The starting node, also known as a root node, contains all the observations in the sample and the internal nodes are repeatedly partitioned until a tree is fully grown and a terminal node is reached. The terminal node represents a point where some stopping criteria has been employed, the node only contains one observation, or all the observations in that node all have the same outcome.

Starting at the root node, each node is split into two groups, called daughter nodes, according to some decision rule that consists of a variable and a split point. Both the right and left daughter nodes are split such that each subset of the data at each node contains a high frequency of the same outcome class, ideally with all the observations in a node of the same outcome class (pure node). Each split is interpreted as a condition set on the selected variable, for example if the decision rule for a particular node is $Age \leq 5$, then any individual that is age 5 or younger is partitioned into one daughter node and everyone older than 5 is in the other daughter node. Each daughter node then goes through the same process of choosing a variable and split that partitions the data into smaller and smaller subsets. Continuing the example described earlier, if the node made up of those older than 5 is followed by a branch that splits the node by gender, then that branch describes a partition that includes females older than five and another partition of males older than five. The recursive framework keeps going until a terminal node is reached. Since each partition is easily interpreted and the overall tree itself is a combination of decision rules describing a specific set of characteristics, these set of rules that can be a very valuable tool in the decision-making process. The hierarchical nature of a tree naturally include interactions that do not need to be known a prior, allowing for very flexible models.

1.1.2 Growing a Tree

The selection of the variable and split utilizes a multi-step selection method that chooses the best possible variable and split among all the predictors in the data. CART is able to identify among the continuous and categorical predictors, for each node, the optimal variable and split point which results in the largest decrease in impurity. The number of possible splits being considered vary by the type of predictor being considered. The possible number of splits for every categorical variable with q categories is $2^(q-1)-1$ [42]. If the variable is continuous with k different observed values, there are k-1 midpoints between the observations to be considered before the decision rule can be determined [13]. To determine

which of these possible splits to select, the splitting rules are ranked and the rule that maximizes the decrease in node impurity is selected.

The purity of each node depends on the proportion of each class in the node. The node is pure when a node is entirely composed of one particular class and most impure when the proportion of each class is equal. The resulting splitting rule maximizes the difference between the node impurity in the parent node and the sum of the impurity in both the right and left daughter nodes [42]. There are multiple measures of the node impurity used to make this determination with a few common methods utilizing either the GINI index or other deviance measurements such as entropy. Once the node is split according to the chosen variable and split-point, each of the resulting daughter nodes are partitioned such that the next best variable and split is chosen from among the remaining predictors not already used in that branch. For example if there are X_1, X_2, X_3, X_4 predictors and some cutpoint in X_1 is chosen at the first split and some cutpoint in X_2 splits the subsequent right daughter node then the choice of X_2 is based off the resulting partition created by the split in X_1 and any other predictor that would have come before it in that branch. The following equations and related notations are adopted from the text in Hastie et. al. [42]. In growing a classification tree, let R_m represent a node m with N_m observations. Let $k(m)$ represent the majority class in node m and P_m represent the proportion of class k observations in node m, with the estimated P_m taking the following formula:

$$\hat{P}_{mk} = (1/N_m) \sum_{x_i \epsilon R_m} I(y_i = k). \tag{1.1}$$

The GINI index and deviance measure impurity in the node by utilizing the proportions of each of the classes in the following formulas, respectively

$$\sum_{k \neq k'} \hat{P}_{mk} \hat{P}_{mk'} = \sum_{k=1}^{K} \hat{P}_{mk}(1 - \hat{P}_{mk}). \tag{1.2}$$

$$-\sum_{k=1}^{K} \hat{P}_{mk} log \hat{P}_{mk}. \tag{1.3}$$

Misclassification rates can also be calculated for each tree to measure overall predictive accuracy of the tree and utilizes the majority voting scheme to determine how often an

observation was incorrectly classified to a particular class[42]. The majority voting takes the class representing the highest percentage in the node and assigns it to every observation in that node using the formula

$$(1/N_m) \sum_{i \epsilon R_m} I(y_i \neq k(m)) = 1 - \hat{P}_{k(m),m}. \tag{1.4}$$

1.1.3 Pruning a Tree

If the resulting tree is very large, the model complexity may over fit the data and if it is too small it may not grasp the important interactions present. Pruning can prevent a model from overfitting the data by removing selected lower branches in the tree. Hastie et. al [42] describe the following cost complexity function, which if minimized, balances the size of the tree with the goodness of fit to the data. This is better than a hard and fast alternative rule for always removing branches that do not achieve some threshold, which can potentially result in interesting relationships being ignored. Since GINI is more sensitive to changes in node probabilities, it is commonly used while growing the tree, whereas the misclassification rates are commonly used for the cost complexity pruning of the tree where the cost function is defined as follows:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T| \tag{1.5}$$

with T indicating any sub tree, $|T|$ denoting the number of terminal nodes in T, α representing the tuning parameter that balances the tree size and its goodness of fit to the data, and Q_m signifying the measures of node impurity. Hastie et. al. describe how to minimize this cost function to produce the optimal tree from the maximal or largest possible tree grown from the data. To minimize the cost function, one must first estimate the tuning parameter α using a 5 or 10 fold cross validation. To do this, either subset the data into 5 or 10 partitions, omitting out one subset at each run on which to test your model. For each run, use the remaining data as a training set to grow the tree and calculate the $|T|$ for the various α values. After each run, use the test data to grow the maximal tree and

test the various values obtained from the test data to determine which α value provides the smallest prediction error. Using these α values, determine the subtree that minimizes the cost function. Subsequently prune the trees so that the values of $|T|$ are reflected and determine which results in the smallest prediction error [42]. In the case where there are missing values and no split point is able to be determined, a surrogate variable may be used. Surrogate variables are predictors and split points that are utilized when there are missing observations in the primary-splitter variable. A surrogate is chosen if they are close to the primary splitter, where the closer they are together, the smaller the information is lost at that split [42].

1.1.4 Variable Importance

Variable importance measurements for the list of parameters are calculated as the decrease in impurity attributed to that variable. The importance of a variable is the sum of the decrease in impurity across all the nodes where that variable was chosen as the primary splitter or as a surrogate splitter[71].

After the tree is grown, the observations in node q are assigned to a class $k(q)$. As a data mining technique, CART is easily utilized for large datasets to classify data into their respective groups. While CART can also handle missing data for its non-parametric and nonlinear categorizations, there are still limitations to this approach [42]. One such limitation is that CART tends to be susceptible to over-parameterization and the resulting prediction can be unstable. This instability is due to the fact that a single tree has high variance, which means that a change in the data may affect which variable is chosen at the first split and then that change is pushed further down the tree [42].

1.2 Random Forests

A single CART tree is known to be unstable, that is small changes in the input data set can lead to difference in prediction. Combining individual models can create ensemble predictors that improve the accuracy and stability of the model. Ensemble predictors, are a supervised learning technique that combines weak classifiers, which can be models that are only slightly better than random guessing, such that the new combined strong classifier has a higher accuracy than each of the individual weaker classifiers[63] .Ensemble predictors display the most improvement when there is high variability between individual models that can be used to smooth out the decisions produced by any single model ([32]).

1.2.1 Bagging

Bootstrap aggregating, also known as bagging, was developed to reduce variance and prevent model over-fitting. Bagging applies a learner to bootstrapped samples of the data, which have been bootstrapped with replacement. If applied to the decision trees, bagging averages the outputs from the repeated bootstrap samples of the data [21]. Averaging produces much more stable estimates, particularly since a bootstrap sample taken with replacement is expected to only have 63.2% of the original data. Through aggregation, Bagging is able to reduce the variance by averaging the prediction over multiple bootstrap samples. In the classification tree setting, B bootstrap samples are selected from the data with B subsequent tree classifiers produced from each tree. The misclassification rate is taken to be the proportion of times the predicted class is different from the true class and averaged across all the classifiers. While bagging works well in increasing precision of highly unstable classifiers, it may potentially result in a worse classifier when the original classifier is already relatively stable[22].

1.3 Random Forest

Random forest is an ensemble classifier developed by Leo Breiman(2001) and is built upon the premise that you can get better performance by de-correlating the trees obtained by Bagging. In Random Forests, B Bootstrap samples are constructed with replacement. To add further randomness and to aid in de-correlating the trees, at each node in building the tree, the best split is chosen from a subset of all possible parameters of size $m$try selected at that node. If $mtry = p$ then, random forest reduces to bagging. Increasing the randomness results in a set of trees that have low correlation and can produce good predictive accuracy and more stable results[23].

It is important to note that each tree contributing to this ensemble classifier is grown without pruning so as to create more variability in the forest. Mtry is an important tuning parameter and locating the optimal $m$try is necessary in finding the best model ([23]). This is especially true in high dimensional problems. The size of $mtry$ must be large enough such that important predictors have a high probability of being chosen within the set.

1.3.1 Random Forest Algorithm

For b = 1 to B:

1. Draw a bootstrap sample of size N from the training data.
2. Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps until the minimum node size n_{min} is reached indicating a terminal node of the tree.
    - Select $m$try variables at random from the p variables.
    - Pick the best variable/split point among the $m$try variables

- Split the node into two daughter nodes

3. Output the ensemble trees T_{b1}^{B}

1.3.2 Out-of-Bag Estimates

Breiman's Random Forest paper includes explanations on how the prediction error is calculated. Bagging is used in the construction of each tree and corresponding classifier. Given a training set T, the classifiers are constructed from the B bootstrap training sets obtained from T and the resulting bagged predictor is calculated from the majority votes. For each X,Y in T, the combined votes over those classifiers which do not contain X,Y are calculated and are called the out-of-bag(OOB) classifier. The generalization error can then be estimated by calculating the error rate for the out-of-bag classifier. In calculating as such, this prevents the need for a testing data set. Since the bootstrap sample excludes approximately one-third of the data, the resulting OOB estimate only includes about one-third of the data points, which can lead to overestimating the current error rate. The OOB error estimate takes the proportion of times that the class with the majority votes is not the true class, averaged over all cases. However, unlike in cross validation, the OOB estimate are unbiased if the estimate is calculated well after the test set error converges[23].

1.3.3 Variable Importance

Variable Importance is estimated through the tree-building procedure. There are two common methods for determining variable importance in random
forests: GINI and permutation importance. The first, which is the default for most statistical packages, is for the split at each node to be determined based on the GINI impurity criteria, where 0 or 1 represent a pure node and 50-50 (in the case of a binary variable) represents the most impure node. The GINI variable importance for variable X_i sums across all trees

the decrease in GINI between the parent and the two daughter nodes for those splits made on X_i [6]. However, utilizing this measure does show a preference for predictors with many possible splits or categories.

This second method is permutation importance, and will be the one that will be focused on in this manuscript. This method uses a permutation framework to calculate the variable importance measurements and changes in prediction accuracy. Calculation of the permutation importance, starts with randomly permuting the OOB cases of X_i in each tree. The other predictors along with this permuted variable are then used for the OOB observations to predict the response. Then the difference in the number of votes for the correct class in the original sample and the permuted sample is computed with the rational being that accuracy will be diminished when an important variable's connection with the outcome is obfuscated. The formula and all related variable definitions for the permutation importance is presented from Strobl et. al's paper on Conditional variable importance for random forests[74] which is illustrated by (equation 1.6). The permutation variable importance measure for X_i is the average permuted difference across all the trees. The idea is that permuting the X_i variable will break the correlation to the response and produce a more accurate measure of variable importance. The permutation importance is more commonly used due to the biased variable preferences that the GINI index tends to produce. Also, when building the forest, the size of the forest can affect the stability of the permutation importance and the larger values of $m$try can result in a large permutation importance [87]. Formally, let

$$VI^{(t)}(X_j) = \frac{\sum_{i \epsilon \bar{\beta}^{(t)}} I(\gamma_i = {\gamma_i}^{(t)})}{|\bar{\beta}^{(t)}|} - \frac{\sum_{i \epsilon \bar{\beta}^{(t)}} I(\gamma_i = {\gamma_{i,\pi_j}}^{(t)})}{|\bar{\beta}^{(t)}|.} \tag{1.6}$$

Equation 1.6 takes $\bar{\beta}^{(t)}$ to be the OOB sample for tree t and ${\gamma_i}^{(t)} = f^{(t)}(X_i)$ is the predicted class for observation i and ${\gamma_{i,\pi_j}}^{(t)} = f^{(t)}(X_{i,\pi_j})$ is the predicted class after permutation. Using this formula the permutation variable importance measure is calculated as

$$VI(X_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(x_j)}{ntree} \tag{1.7}$$

This calculation for variable importance looks at the magnitude of the difference between the original prediction error and the prediction error once the values for that variable are permuted [75]. The stronger a signal or the more informative that variable is, the higher in the tree that variable will appear. Since the subsequent branches are conditional on the variables selected higher in the tree, any changes to important variables at the top of the tree can result in a larger effect on the overall prediction accuracy of the tree.

The variable importance averaged across the trees can also be standardized to be used in hypothesis testing. This overall variable importance measure for each variable can be transformed into a scaled variable importance measure by dividing the variable importance by the standard error $\hat{\sigma}/\sqrt{(ntree)}$ [75]. This scaled variable importance measure now has a standard normal distribution which can be used in hypothesis testing to determine if the variable importance score is significant at an alpha level.

1.3.4 Proximities and Applications

A useful measure that is calculated from the random forest algorithm is the proximity matrix. The proximity matrix indicates how close two observations are to each other. The (i,j) element in the proximity matrix is the average number of trees that have observations i and j end up in the same terminal node [48].

Random forest also has methods to deal with potentially large amounts of missing data. In Breiman's seminal paper([23]), he proposed a rough fix to deal with missing values in the data. The proposed method is simplistic and substitutes the median value for the continuous variables and takes the class with the majority votes for the imputed categorical value. Comparatively, a more adaptive method could utilize the proximity matrix that

is already calculated from random forest. To take advantage of this calculation, random forest package in R [61] is grown using the data with the missing values filled in using the simplistic method. In the package, the rough imputed values are then updated using the proximity matrix as weights. For continuous variables, the R function updates the imputed values using the matrix as the weights and taking the weighted average of the non-missing observations. Similarly, the package takes the categorical variables take the maximum value from the proximity matrix that corresponds to that variable.

Along with handling missing data, the proximity matrix is useful in determining outliers. For a specified class, low proximities relative to the others in that class, may be indicative of an outlier [6].

1.3.5 GINI bias

Even though CART and other decision tree methods commonly use the GINI index to determine the best variable and cut-point to use in growing the tree, Strobl's 2007 paper [73] indicates that the GINI measurement can be biased. In fact, Strobl indicates that the estimate of GINI does not produce an unbiased estimator and results in preferences for correlated and/or continuous predictors. If N represents the sample size and p corresponding to the proportion of the majority class in that node, the estimate of impurity, namely the GINI index, underestimates the true GINI index by a factor of $(N-1)/N$. Also, the paper indicates that the expected change in GINI($\hat{\Delta G}$) between the parent and daughter nodes is equal to $2p(1-p)/N$. Therefore, under the null hypothesis that the change in GINI is 0, $\hat{\Delta G}$ has a positive bias that is a function of the sample size. When the predictors have different sample sizes, there is a bias towards those with many missing values (smaller number of non-missing observations). Strobl's paper also indicates that by testing multiple cutpoints to find the optimal split, a multiple testing situation occurs that serves to increase the type I error rate. For the splitting selection situation, the type I error is when a variable is chosen for splitting even if it is just a noise variable. In the decision tree situation with binary

splits, the number of comparisons that need to be performed to determine the optimal split depends on the number of possible splits of that variable. This means that as a result of all the multiple comparisons, variables with many categories or continuous variables twill be chosen more frequently when the GINI index is used to measure impurity [73]. The next section will deal with this issue.

1.4 Conditional Inference Forests (CIF)

It is known that Random Forest variable importance is biased towards correlated predictors [75]. Highly correlated variables arise in many situations, especially in biologic and genomic studies where variables can be subsets or even linear combinations of variables. Biologic and genomic studies often have many more predictors than observations, thus machine-learning techniques such as Random Forests are often used for feature selection. In Strobl's paper [75], a conditional variable importance measure was proposed to address this limitation and reduce biased variable importance measures produced by the Random Forest algorithm. Their main focus was on distinguishing between the marginal associations that produce relatively high variable importance and the more informative associations present once conditioning on other predictors. Strobl's method builds upon the work done to produce CARTscans plots that output marginal influence along with conditional influence plots for categorical predictors. The rationale for conditional variable importance measures stem from the fact that both the permutation importance and the GINI importance measures indicate marginal associations and thus may be misleading. By conditioning on other predictors in the data, insights into the real relationship between the variable of interest and the outcome will be availed. By conditioning on other predictors in the data, the variable importance measure can break free from the preference of correlated predictors to be chosen in early splits, which translates into a higher variable importance score. The situation using decision trees lends itself nicely towards a conditional framework. Since at any node, the prior splits in a branch describe a pattern of predictors and splitting criteria, that node can be thought

of as conditional on only those particular variables seen prior to it in its branch. The only exception is at the first split, when there is no other predictors prior to it to condition on [75].

1.4.1 Conditional Inference Trees

The basis upon which conditional permutation Random Forests is built hinges on the unbiased tree method proposed by Hothorn, Hornik and Zeileis in 2006 [44]. Their method proposes a two-step procedure that separates the variable selection from the cutpoint selection used to partition a node into it's two daughter nodes. Hothorn's paper illustrates that separation of the variable selection and the cutpoint determination does well in preventing the tree algorithm from preferentially choosing categorical variables with many categories or variables with many missing values. Their paper describes how using hypothesis tests to determine the stopping criteria allows the predictive accuracy to be the same as that from an optimally pruned tree while preventing overfitting. However this aspect of the separation in determining the decision rule is not as important in this situation since averaging the trees takes care of any overfitting that any one individual tree may have[44].

The following description of the steps utilized to determine what variable and split point to use at each node is presented from Hothorn's paper [44]. A two-step method is described with the variable selection step testing the global hypothesis of independence between any of the covariates and Y. The hypothesis is made up of p partial hypotheses that test the hypothesis that the distribution of Y given the covariate is the same as the marginal distribution of Y, with the global hypothesis looking at all of the covariates. The association is measured by permuting the responses and a function of the covariate being tested. The p-value from each test is obtained using the permutation framework that fixes the covariates and conditioning on all the possible permutations of the outcome. If the global hypothesis fails to be rejected at some α level, then the node is not split and becomes a terminal node. If it can be rejected, then the covariate with the strongest association with

the outcome is chosen to split the node. Once the variable is chosen, the next step determines at what cutpoint to split the variable. Once the best predictor is selected, then the cutpoint selection is selected by again doing a permutation test. The potential cutpoint splits the observations in the node to form two subsets of the observations for that node. The split that corresponds to the case where the permutation test maximizes the difference between the prediction permuted for each subset of the node, is the one chosen for the splitting criteria [44].

1.4.2 Cforests

Strobl et. al.'s 2008 paper on conditional variable importance for random forests[75] goes into detail regarding the rationale behind the conditional framework for the permutation variable importance. Since permutation variable importance may lead to spurious associations, consider it in terms of permutation tests. With a global test, the null hypothesis is that Y is independent from all the other predictor variables in the data. This translates into the null hypothesis that the permutation of Y has no effect on f(y) or on $f(X_1, X_2, ...X_p)$. This implies that when the permutation importance shows a significant difference in the joint distribution of Y and $X_1, ...X_p$, that either X_j is not independent of Y or that the X_j is not independent of the other p-1 covariates. In determining how influential X_j is on Y, the relationship between X_j and the other p-1 covariates are not of interest. To achieve this, a conditional permutation scheme only looks at if X_j and Y are independent instead of if X_j is independent of both Y and the p-1 covariates. This is done by permuting X_j only among the observations that have the same values in each of the p-1 covariates to maintain the same correlation structure among the predictors. The null hypothesis now is that X_j is independent of Y given $X_1, ...X_{j-1}, X_{j+1}, ...X_p$. If we let $Z = X_1, ...X_{j-1}, X_{j+1}, ...X_p$ then $H_o : (X_j \perp Y)|Z$. Under this null hypothesis the following conditional results.

$$P(Y|X_j, Z) = P(Y|Z) \tag{1.8}$$

If X_j is correlated with Z, then the differences in the distributions are a result of such correlation and not the relationship of interest and thus leads to the preference in the variable selection and overestimates the permutation importance. Also by permuting within groups of observations with $Z = z$, this maintains the correlation structure of X_j and the other covariates. The resulting unbiased conditional permutation trees are used to define the partitions upon which the permutations will be done. This provides a readily available partition grid that has already been determined by the algorithm[75].

1.4.3 Cforest Algorithm

The following algorithm is presented from Strobl's paper [75].

1. For each conditional permutation tree, compute the OOB prediction accuracy before the permutation of X_i.
$$\frac{\sum_{i\epsilon\bar{\beta}^{(t)}} I(\gamma_i = \hat{\gamma}_i^{(t)})}{|\bar{\beta}^{(t)}|} \tag{1.9}$$

2. For each of the variables Z to be conditioned on, obtain the cut points that split the variable in the current tree. Then obtain the partition by using each of the cut points sequentially.

3. Within the partition grid, permute the values of X_i and determine the OOB prediction accuracy using the following formula:
$$\frac{\sum_{i\epsilon\bar{\beta}^{(t)}} I(\gamma_i = \hat{\gamma}_{i,\pi_j|Z}^{(t)})}{|\bar{\beta}^{(t)}|}, \tag{1.10}$$
where $\hat{\gamma}_{i,\pi_j|Z}^{(t)} = f^{(t)}(X_{i,\pi_j|Z})$ is the class that was predicted after permuting X_j over the partition structure.

4. The average difference between the prediction accuracy of the unpermuted and the permuted grid, computed across all the trees, results in the permutation variable importance.

1.4.4 Partition Grid

To determine what variables to condition on, Strobl's paper recognizes that a natural choice would be to take advantage of the model already created from each of the individual conditional inference trees grown in the forest. The paper proposes that each of the trees can be thought of as a series of binary splits that produces at least one cutpoint for each variable that can be used of as a partition grid. However since each of these cutpoints are based on nodes within a branch, Strobl notes that bisecting that variable does not necessarily split the sample space but can result in partial planes and can make things too computationally intensive. Instead of using a cutpoint specified at each node, Strobl suggests using the cutpoint, regardless of whether it is based on a continuous or categorical predictor, as a bisector of the sample space to produce a simpler grid. Also as a way to alleviate some of the computational burden, Strobl proposes only using those predictors correlated above a certain threshold with the variable of interest to be conditioned on. The partitioning method could result in a grid that contains some small cell frequencies. Strobl indicates that while this is was not a problem in her simulation studies, the small cell frequencies only add to inducing greater variation for the resulting ensemble predictor[75].

One caveat to this method is that $m$try has a strong influence over the effectiveness of this variable importance measure. Selecting an $m$try of 1 results in a random variable selection regardless of whether the permutation or conditional importance measures were used [75]. On the other hand, a high $m$try will allow a conditional permutation to have the most effect at preventing spurious relationships, it will also result high variability in the values of the importance measurement Strobl [72]. But regardless of what $m$try is chosen, the conditional permutation framework results in a similar pattern but with a lower variability than the permutation importance, which in turn may help with identifiabilityStrobl [72]. The biggest caveat with this method is its computational intensiveness. It is not feasible for even moderate sized datasets.

1.4.5 Variable Importance Bias

Strobl's 2008 paper also acknowledges that while the conditional permutation results in less biased importance scores as opposed to using the GINI index, the conditional variable importance algorithm still resulted in the uncorrelated predictor variables being selected less often and with a lower importance in the hierarchy of the tree and thus resulting in a low variable importance measure. The tuning parameter $m$try, as mentioned before, is also highly influential, with a low value preferentially choosing correlated predictors and a large value increasing the variability of the importance measure.

Another study done by Nicodemus and Malley(2009) also aimed to deal with the issue of variable importance bias. Their study found that Random Forest preferentially selected correlated predictors when the selection was obtained using the GINI index and conditional inference forests tended to overweight the uncorrelated variables. They also found that conditional inference forests were computationally infeasible for moderate to large datasets [53].

1.5 Fuzzy Forest

Fuzzy Forests and all related concepts described below are presented by Conn et. al.'s paper Fuzzy Forests: Extending Random Forests Algorithm for Correlated, High-Dimensional Data [28]. Fuzzy Forests is a screening algorithm to find the most important variables when there exists correlated variables along with independent variables, especially when the number of parameters greatly exceeds the number of observations. Fuzzy Forests is a two-step process that utilized both unsupervised and supervised learning. The process starts with unsupervised learning where a weighted correlation network separates the feature space into modules, such that only the features within a module are highly correlated and there is low correlation between modules [28].

1.5.1 Module Selection

The first step is constructing a weighted correlation network construction or module creation. The discussion requires that some key network concepts be addressed. The first concept that needs to be explained is "'approximately scale-free"' networks. Barabasi and Bonabeau[17] explain that some networks, such as those modeling relationships between sexual partners, portray a network that contains some individuals with few partners and others that are hubs with hundreds of partners, are called "'scale-free"'. "'Scale-free"' is used to loosely describe some hubs' ability to have seemingly endless links and no node is typical of the others. In the past, complex networks were thought to be completely random and are characterized by all nodes having approximately the same number of links. In this setting, the distribution of the number of connections for each node follows a Poisson distribution. However, in many real life networks, the "'scale-free"' setting, is more appropriate. These networks indicate that the probability that a node has k links follows a power law distribution and is proportional to $1/k$. Unlike the Poisson distribution used to describe random networks, which does not allow for hubs, the power law for the "'scale-free"' networks allows for networks in which a few hubs dominate. In many real life situations, the random networks ignore hubs and fail to describe what is truly occurring due to the underlying assumption that all nodes are equal and existed before the links were made. However, in real life networks that are constantly evolving, older nodes have greater chances to gain more links and preferences are placed on certain nodes(ex. people are only familiar with a small portion of the internet and choose from a tiny subset of the more popular sites because they are easier to locate) [17]. Another simple example are airline networks where most flights originate from several hubs such as LAX, JFK, DFW etc.

Formally, Barabasi describes scale-free networks [60] as any network where the probability of a node having k links to others in the network follows a power law distribution which is $Ck^{-\gamma}$, with C being a constant and γ is the degree exponent. Barabasi indicates that since the power law diverges when there are no links to other nodes, k_{min} represents the

smallest k for which the power law holds. When k is continuous, the normalizing constant C is represented by $(\gamma - 1)k_{min}^{\gamma-1}$. When k is continuous, $p(k)$ does not indicate the probability that a randomly chosen node has degree k like it does when k is discrete, instead the probability can only be defined within a range of k_1 to k_2. In this case, the probability that a node has a γ between k_1 and k_2 is defined as $\int_{k_1}^{k_2} p(k)dk = (\gamma - 1)k_{min}^{\gamma-1} \int_{k_1}^{k_2} k^{-\gamma}dk$.

Barabasi also indicates that the natural cutoff or the size of the largest hub is calculated as

$$k_{max} = k_{min} n^{\frac{1}{\gamma-1}} \tag{1.11}$$

where n is the total number of nodes, and indicates that the larger the network, the larger the hubs become. In most real networks, γ is ≥ 2 since $\frac{1}{\gamma-1} > 1$ in the k_{max} formula when $\gamma < 2$. If this were the case, this would mean that the number of connections to the largest hub would grow faster than the size of the network. γ is usually between 2 and 3. γ in this interval only have finite first degree moments and all higher order moments diverge as $\lim_{n\to\infty}$. When γ exceeds 3, these networks resemble random networks. As a result, if $\gamma > 3$ then the required network size needed for a scale-free network is a transformation of (1.11) to $n = (\frac{k_{max}}{k_{min}^{\gamma-1}}$. For example if $\gamma = 5$, $k_{min} \approx 1$, and $k_{max} \approx 100$ then a scale-free network would require at minimum 10^8 nodes, which most networks are not [60]. Zhang and Horvath [86] proposed an easier method using R^2, which is the square of the correlation between log(p(k)) and log(k), to determine if a network approximately has scale-free topology. These network characteristics along with other network concepts will help to determine the groups of variables or modules in the network. Zhang and Horvath use their criteria for approximate scale-free topology along with other network concepts to present an overview of module selection. Zhang and Horvath describes the network topology as a graphical representation of a network, where the vertices are the variable and the edges are the interactions between them. Two variables are connected in a co-expression network if the co-expression, measured using some measure of similarity such as Pearson's correlation coefficient, is above some threshold. The co-expression network links to an adjacency matrix, which indicates the connection strength between variables. The correlations between corresponding variables are calculated as a

similarity matrix, which is then altered into the adjacency matrix. The adjacency matrix is then used to represent node dissimilarity, which will serve as the input data upon which the variables will be clustered into modules or clusters of variables [86].

Zhang and Horvath describe an adjacency function which uses Pearson's correlations from the similarity matrix to create the adjacency matrix. The adjacency function can utilize either a strict cut-off, which is known as a hard threshold, or the more flexible, soft threshold. The most common adjacency function for hard thresholding is

$$a_{ij} = \begin{cases} 1 & \text{if} s_i j = cor(i,j) \geq \tau \\ 0 & \text{if} s_i j = cor(i,j) < \tau \end{cases}$$

with τ being the hard threshold. Some suggestions for the estimation of τ utilize the significance level of a correlation, which in turn thresholds the p-value corresponding to the correlation coefficient. The size of the network decreases as a function of the threshold [86]. An alternative method uses the relationship between the network size and the correlation threshold and sets the network size as a constant [19]. Hard thresholding has an intuitive interpretation since it represents the number of variables that are directly connected with an edge(first order interaction), but the lack of flexibility does not recognize those connections that are close to the threshold cutoff [86].

Zhang and Horvath's paper describes a soft thresholding function such as the power adjacency function, is given as $a_{ij} =\mid s_{ij} \mid^{\beta}$, where β is the soft threshold . The power adjacency function can be used to express a weighted correlation network. It also has a factorization property that preserves the factorization of the correlations such that if $s_{ij} = s_i s_j$ then $a_{ij} = a_i a_j$ with $a_i = (s_i)^{\beta}$. Estimating the β parameter in the soft-threshold is different from that for the hard-thresholding. In this case, the soft-thresholding parameter uses the scale-free topology criteria, which only considers networks with $R^2 > .8$. Since β values that result in a R^2 value close to 1 may show networks with very few connections. Additional considerations should include a high mean connectivity to ensure enough information for module creation, and that the slope of regression line between log(p(k)) and log(k) should

be close to -1.

Following the adjacency matrix calculation Zhang and Horvath describe how the topological overlap measure is computed. They measure the overlap in the variables with strong connections to both variables i and j. These values populate the topological overlap matrix (TOM) and is calculated as

$$w_{ij} = \begin{cases} \frac{l_{ij}+a_{ij}}{min(k_i,k_j)+1-a_{ij}} & \text{if } i \neq j \\ 1 & \text{else} \end{cases}$$

In this formula, $l_{ij} = \sum_u a_{iu}a_{uj}$ and node connectivity is represented by $k_i = \sum_u a_{iu}$, which represents the variable connectivity by taking the sum of the ith row in the adjacency matrix. Under the hard thresholding, l_{ij} represents the number of variables that the ith and jth variables are both highly correlated with. The topological overlap w_{ij} is 0 when there is no variables common to both the ith and jth variables. Similarly, w_{ij} is 1 when all of the neighbors of the ith node is the same as those of the jth node. As a result of how the w_{ij} is calculated, the TOM is non-negative and symmetric. The TOM is then transformed into a dissimilarity matrix by taking each element in the matrix and subtracting it from 1. The TOM dissimilarity matrix is then used to determine the modules [86].

Once the TOM dissimilarity are computed, the modules with high topological overlap can be determined [86]. The modules are determined using average linkage hierarchical clustering. The following description by Sayad, describes hierarchical clustering creating clusters based on prior clusters ([2]). Hierarchical clustering groups all the variables divisively by starting with one cluster and then recursively partitions the cluster into two groups that are the most dissimilar, until there are n clusters. The agglomerative method is another hierarchical clustering method where each node is its own cluster. Then a similarity measure is computed between each cluster and similar clusters are combined. Successive clusters are combined until a single cluster is formed. Horvath's book on weighted network analysis

describes clusters being combined using the pairwise dissimilarity measure [43]. The linkage method or inter-cluster dissimilarity is calculated between each pairwise nodes in the clusters. Horvath's equation on average linkage hierarchical clustering is computed as

$$d_{average}(clust.q1, clust.q2) = \frac{\sum_{i \in clust.q1} \sum_{j \in clust.q2} d_{i,j}}{\mid clust.q1 \mid\mid clust.q2 \mid} \tag{1.12}$$

$d_{i,j}$ represents the pairwise dissimilarity between the ith and jth node and $\mid clust.q1 \mid$ and $\mid clust.q2 \mid$ is the number of objects in the two clusters. For illustrative purposes, Horvath indicates that the results of agglomerative hierarchical clustering as being represented in a dendrogram, with nodes (x-axis) combined at each step and the height(y-axis) of each step indicates the dissimilarity of the merged nodes. The number of merges is less than $n - 1$, with the heights at each merge increasing. The resulting plot resembles a tree where the branches correspond to clusters and nodes as leaves. The nodes and corresponding branches are organized such that the lines in the dendrogram do not cross [43].

Once the dendrogram is created, Zhang and Horvath [86] describes the module creation where branches in the dendrogram depict the different modules. The TOM plot uses the TOM-based dissimilarity values to create a dendrogram. A Topological Overlap Matrix Plot sorts the nodes by the hierarchical clustering tree and represents the TOM dissimilarity values utilizing complementary colors. Since the TOM-based dissimilarity matrix is symmetric, as is the TOM plot, the modules are represented by high overlap, which is found along the diagonal [86].

1.5.2 Recursive Feature Elimination with Random Forests

It is important to note that the weighted correlation network in the first step is done without the outcome variable, that is, it is unsupervised learning. Once the modules are formed, variable screening using recursive feature elimination on each module is used to reduce the parameter space of each module.

Diaz-Uriarte and Alvarez de Andres [31] proposed an iterative Random Forest method. Their method entails iteratively building Random Forests and removing those variables with the smallest variable importance, such that the resulting subset of variables produces the smallest OOB error rate. For this procedure, the OOB error is used solely to select the final set of variables and not for estimation purposes, since the usage of the OOB in the iterative approach results in the OOB being biased down. This is similar to the rationale for the "'selection bias"' explained by Ambroise and McLachlan [14]. In it, they describe that a cross validation or the bootstrap method[33] should be utilized to correct for the selection bias resulting from the feature-selection process.

The selection bias described by Ambroise and McLachlan [14] arises from which variables are used to perform feature selection in the training of the classifier. When the data is partitioned into a training and test set, a selection bias results due to the fact that the test error is based on the test set, which is a subset of all the variables used to create the classifier. The end result is that the test error is larger than the prediction error. For example, if three genes are selected, they report that the data is split into a 95/5% for the training and test sets, respectively. Fisher's linear discriminant rule had an average test error or 10.7 and 0% [14].

Fisher's linear discriminant function classifies the data into two groups, where each group are classified based on k variables. The data is transformed into univariate observations such that the data from each group are separated as much as possible[84].

The bootstrap method described by Efron and Tibshirani [33] uses the bootstrap error, B1, which predicts the error at x_j from only the bootstrap samples that do not contain x_j. The bootstrap sample contains 63.2% of the original data. The $B.632$ estimator corrects for this bias by taking the weighted average of the bootstrap error and the training data error rate (resubstitution error) [33].

Diaz-Uriarte and Alvarez de Andres utilize the bootstrap method above, also called

.632+ bootstrap method, to determine the prediction error rate, since the weight used in the bootstrap method reflects the amount of overfitting. The bootstrap method is used on the complete procedure which uses samples not selected in the Random Forest or variable selection method to compute the bootstrap error. They go on to describe their method which looks at all forests that are produced from iteratively removing the least important variables. The default is that the lowest 20% are removed. However, this tuning parameter can be adjusted depending on the resolution needed. Lowering the resolution, requires that more variables are removed at each iteration and which speeds up the algorithm. After fitting all the forests, the OOB error rates are compared and the forest producing the smallest number of variables whose error rate is within u standard errors of the forest with the lowest error rate. This can lead to the selection of a smaller subset of variables which would have produced similar error rates [31].

1.6 Fuzzy Forest

Fuzzy Forest described in the paper by Conn et. al. [28] utilizes both the feature selection procedure described by Diaz-Uriarte and Alvarez de Andres and the module clustering described by Zhang and Horvath. After the predictor variables are separated into modules based on the weighted correlation network, the important variables are distilled down to a small subset of predictors by iteratively performing feature space and feature elimination Random Forests (RFE) on each of the modules.

Specifically, within each module, the RFE random forest is performed with the least important features being removed after each run, with the process only stopping when a user specified minimum number of predictors are retained. Once the set of reduced predictors from each of the modules are obtained, RFE is again performed using all the important predictors from each module that are still retained, the "survivors". This results in the final list of important predictors. The user specifies the number of variables they wish the Fuzzy

Forest to choose and that is what is returned.

Fuzzy Forest reduces the computational cost since only smaller subsets of the predictors are used each time Random Forest is built and thus is suitable for large datasets. However, since Fuzzy Forests builds the Random Forests using only variables in the same module and therefore have a similar correlation structure, uncorrelated predictors are no longer in competition with correlated variables as they are in their own module (the grey module by default), resulting is a decrease in the bias towards correlated predictors. Similarly, when the important features from each module are combined in the final overall RFE-RF, this allows for interactions between modules [28].

Steps of the algorithm:

1. Create Modules
   (a) calculate correlation between each pair of predictors and raise to the power β
   (b) transform the similarity matrix into an adjacency matrix that measure connection strength
   (c) use the adjacency matrix to calculate the topological overlap matrix
   (d) transform TOM into the dissimilarity matrix used
   (e) use the dissimilarity matrix to combine nodes based on average linkage hierarchical clustering and form the modules
2. Feature Selection Random Forest on each Module
   (a) for each module i:
      - perform random forest on that subset of the predictors and remove the lowest $k\%$ of the predictors from the variable importance list.
      - repeat the random forest on the reduced subset of predictors from the previous step. Continue until the selected subset of variables produces the smallest OOB error rate.

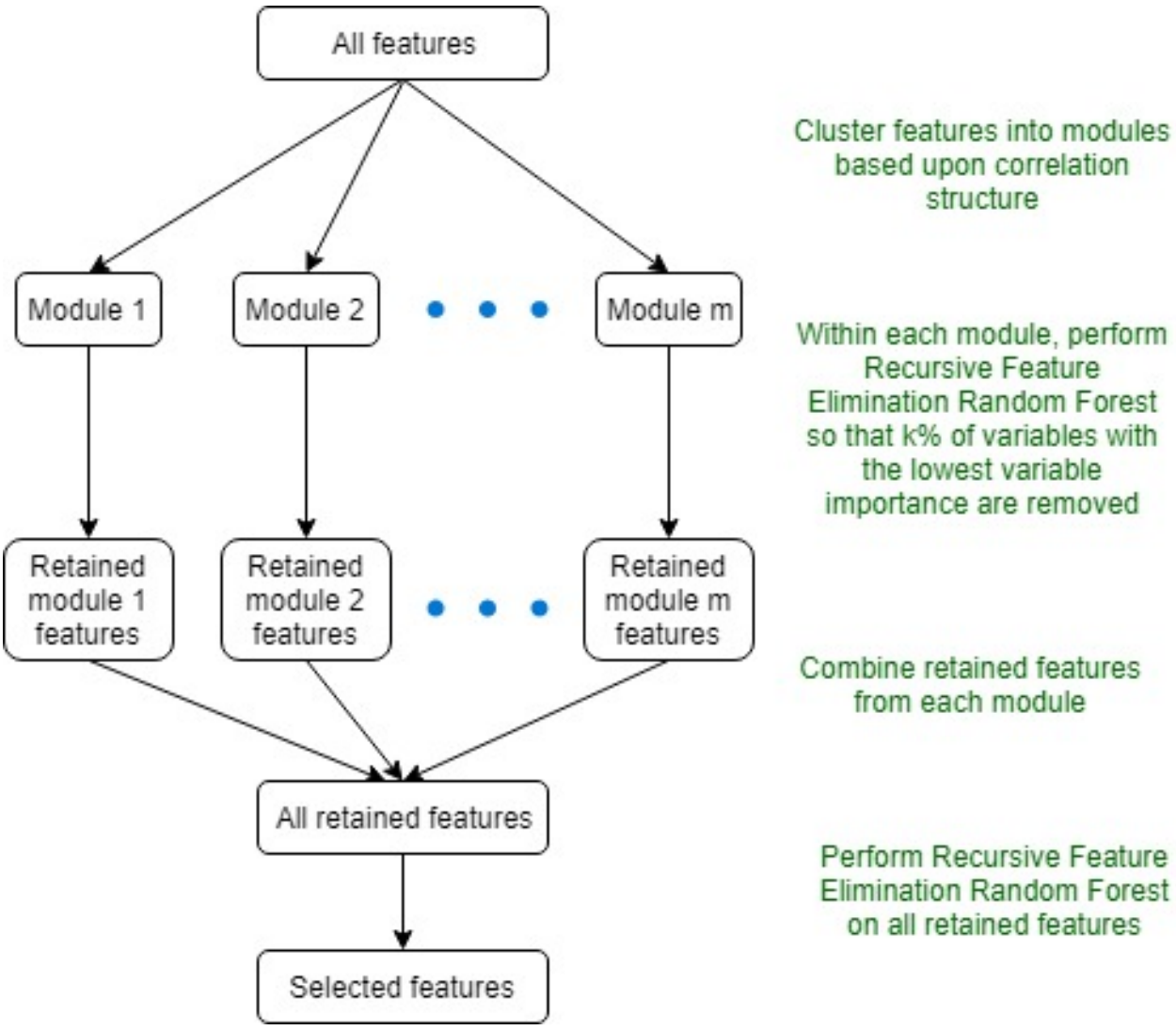

Figure 1.1: Fuzzy Forest Algorithm

3. Feature Selection Random Forest selecting from all "Important" Variables

   (a) Perform RFE again using the final set of predictors that resulted from step 2.

CHAPTER 2

Applications to nursing and genomic data

The decision tree methods previously discussed are useful for analyzing complex and high-dimensional datasets, which can be difficult to work with when analyzed using parametric regression methods commonly used by many researchers. They provide useful alternative methods that can explore the full range of variables available. Nursing research is a field where on many occasions, multiple questionnaires are employed to garner demographic characteristics and other relevant information based on the theory of the relationship being explored. While there is no limit on the quantity of questions being asked, budget constraints commonly limit the number of individuals' from whom information being collected. It is not uncommon for the size of a study to include less than thirty subjects while the quantity of variables collected can far exceed that number. This is also commonly the case for genomic data which, if collected on humans, can contain information from about 3 billion bases [8]. In both cases, the number of variables collected can far exceed the number of subjects in the data. The following chapters explore various contexts that utilize decision tree methods, Fuzzy Forests in particular, to explore the relationships between the predictor variables and the outcome.

2.1 Application 1: Recidivism Among Homeless Men

California has high recidivism rates, which have been estimated to be larger than 50% [12]. Identifying factors that influence an individual being re-arrested could have a huge public health the impact on not only the lives of these individuals but on overcrowding in jails and prisons in California. Recently released individuals from prison and jail are prone to numerous hardships that are impediments to re-acclimating to life outside of incarceration. Hardships including limited access to employment, adequate housing, treatment for drug addiction ([40], [80]), along with substance use and abuse ([58],[59]). Many former prisoners also face homelessness [58]. Indeed many inmates have struggles with mental health disorders, previous imprisonment, substance abuse, and poor health status [36] and are associated with homelessness prior to incarceration.

Nursing interventions have been created to study this problem. One such study of homeless men which was collected by Nyamathi et al. [56] utilized nine questionnaires and collected additional information that includes demographic characteristics, sexual behavior, criminal history, general health, and family history. The study was a randomized controlled trial that studied three alternative interventions offered to former inmates in Southern California who were residents of a residential drug treatment program and were homeless at the time of their release from jail/prison between February 2010 to January 2013: (1) peer-coach and nurse case managed (PC-NCM) program; (2) peer coach (PC) program with brief nurse counseling; and (3) a usual care (UC) program with brief PC and brief nurse counseling. The data collected from this study will be explored further utilizing decision tree methods to identify which among the 255 variables from 534 male ex-offenders are predictive of re-arrest within 6- or 12-months post release.

2.2 Application 2: Lithium Response Among Bipolar Individuals

Similar to the situation with the homeless male ex-offenders, decision tree methods can also be useful in identifying Single Nucleotide Polymorphism (SNPs) related to lithium response in those with bipolar disorder. Bipolar disorder is a mental disorder affecting mood and energy/activity levels which are punctuated by recurring episodes of "highs" and "lows". The severity of symptoms varies by person, and the average onset of bipolar is at age 25. Each year, 2.6% of U.S. adults are diagnosed as bipolar with 82.9% classified as severe bipolar [10]. Lithium is a mood stabilizer and is a primary treatment for bipolar disorder. While many patients respond to lithium, approximately 30% of patients are non-responders or partial responders [34]. There is inconclusive evidence of a genetic component to bipolar disorder ([30], [29]). Identifying a genetic link between bipolar disorder and lithium response could allow for effective treatment for those suffering bipolar disorder that would respond to lithium treatment and prevent the unnecessary treatment of patients who will not respond to this form of treatment.

The bipolar data analyzed was collected by the Genetic Association Information Network (GAIN), an NIH funded a study of bipolar disorder. Cases and controls were genotyped with a Translational Genomics (TGEN) sample being a subset of this data. The TGEN sample contains 1190 genotyped bipolar disorder cases from the Bipolar Genome Study along with 401 controls. The sample collection from the GWA study included genotyping using the affymetrix genome-wide human SNP array 6.0 [37]. A drug questionnaire was also collected from these subjects and included information on their lithium response. Two subsets of the data, which consist of a subset which contained information on lithium response, were used as a separate training and test set and were analyzed using Fuzzy Forests.

2.3 Application 3: Willingness to Quit Among Smokers Who are Nurses and Health Care Professionals

Smoking has a dramatic effect on Public Health. Health care professionals are uniquely informed as to the dangers of smoking. While exploring factors associated with smoking is not a novel undertaking, perhaps Fuzzy Forests can provide new insights into rationales behind smoking in the health care worker population. To this end, the 2010-2011 Tobacco Use Supplement of the Current Population Survey (TUS-CPS) data is analyzed. Among the subset of health care professionals, the focus is on current everyday smokers. We assessed subjects who are interested in quitting smoking and have taken active steps to stop (Tryhards) and also those that have stated that they are not interested in quitting smoking (Diehards).

The TUS-CPS uses stratified probability sampling to provide representative estimates of the population by occupation and has been administered since 1992 with data being collected every 3-4 years [4]. The data was subsetted to include only those in health care related occupations, such as dentists, pharmacists, nurses, therapists for example, that are current everyday smokers. The final dataset included 876 individuals with 99 potential covariates being retained from the original set of predictors. The goal of the analysis was to find predictors to determine if a person is likely to be a Diehard or a Tryhard. A Diehard is anyone who indicated that 1) they had not stopped smoking for one day or longer in the past 12 months because they were trying to quit, 2) had never made a serious attempt to stop smoking even for a day, and 3) that they did not indicate that they are seriously considering quitting smoking within the next 6 months, and 4) that their score for interest in quitting was below 7 in a scale of 1 to 10. A Tryhard is a non-Diehard individual that is also very interested in quitting smoking, indicating an 8 or higher out of 10 in their interest scale of quitting smoking.

2.4 Analysis

These three different applications are explored further in the following chapters. In these chapters, various decision tree methods, such as those described in the previous chapter are compared. Exploration of the homeless male ex-offenders utilizes CART, Random Forests, Fuzzy Forests, a modified version of logistic regression along with a penalized regression model, LASSO. The bipolar dataset, exploring the genetic component of lithium responders, is analyzed using Fuzzy Forests and logistic regression. Lastly, the Diehards and Tryhards are explored using unweighted Fuzzy Forests and weighted Fuzzy Forests followed by a weighted logistic regression using the variables selected via Fuzzy Forests.

CHAPTER 3

Exploring Factors Associated with Re-arrest among Homeless Adults Using Statistical Machine Learning Techniques

3.1 Abstract

Background: Homeless adults are at high risk for re-arrest within 6 to 12 months of release. The present study compares alternative statistical prediction methods, including machine-learning techniques, in the context of a study evaluating a nursing intervention that aimed to guard against re-arrest for recently released homeless offenders. The study collected data on a multitude of factors such as subjects demographic characteristics, social support, judicial involvement, physical health, and mental/emotional health.

Purpose: This paper presents a recently developed machine learning technique called Fuzzy Forests and compares its performance to other existing methods such as classification and regression trees (CART) and Random Forests in identifying important variables for modeling re-arrest among homeless adults.

Methods: The variables in the analyses included demographic characteristics, childhood and family background, peer relations, knowledge and attitudes about hepatitis, various tools assessing mental, emotional and physical health, drug and treatment history, sexual behaviors and criminal history. Predictors from decision tree methods, specifically CART, Random Forests and Fuzzy Forests are compared to predictions from a parametric logistic

regression model and from a semi-parametric group LASSO procedure. Within each approach, the 15 most important variables (out of 210) related to re-arrests are identified. Due to quasi-complete separation in the data when all of the predictors were considered simultaneously, logistic regression was implemented using a stepwise procedure within modules identified by the Fuzzy Forests procedure. Area under the curve (AUC), which can be interpreted as the probability of correct ranking, along with the misclassification rate are two objective measures that are used to compare the performances of these methods.

Results: The five methods identified several common variables in their list of the 15 most important variables for re-arrest. The resulting models applied to a hold-out sample indicate that Fuzzy Forests has the lowest misclassification rate and the highest AUC out of these five methods.

Conclusion: Machine learning methods can be very useful in studies with large number of variables by efficient dimension reduction, thereby facilitating comprehensive predictive modeling. A new method called Fuzzy Forests is useful when some of the explanatory variables are correlated, as is likely among items within an instrument such as the CES-D. We found that Random Forests, Fuzzy Forests, CART, and stepwise logistic regression within modules and group LASSO have comparable AUCs. However, the results from the using the hold-out sample as a test set shows that Fuzzy Forests produces the highest predictive accuracy. Overall, we conclude that the decision-tree machine learning techniques are comparable in dimension reduction, with the Fuzzy Forests technique having a theoretical appeal in its capacity to deal with correlated predictor variables while logistic regression needed input by the analyst to avoid computational problems and the group LASSO procedure yielded mixed results.

3.2 Introduction

According to the California Department of Corrections and Rehabilitation, among adults who were convicted of a felony and incarcerated and then released in the 2012-2013

fiscal year, the one-year recidivism rate for arrests has been estimated to be greater than 50% [12]. This includes only those who return to state prison. These individuals face inherent difficulties in adjusting to life outside of prison, including limited access to employment, adequate housing, and treatment for drug addiction ([40], [80]). Substance use and abuse remains a substantial problem in this population ([58],[59]) . Many also face homelessness, with the lack of suitable housing and access to rehabilitative programs making reentry into society upon release from prison difficult [58]. In particular, mental disorders, previous imprisonment, substance abuse, along with poor health status [36] were found to be associated with homelessness prior to incarceration among prison inmates. Identifying the factors that contribute to recidivism has the potential to aid legislators and social workers in allocating limited funding to help service providers to address the needs of former inmates in ways that protect against re-arrest.

In an effort to address these issues, a recent randomized controlled trial (RCT) was conducted by Nyamathi et al. [56] studying three alternative interventions offered to former inmates in Southern California who were residents of a residential drug treatment program who were homeless at the time of their release from jail/prison between February 2010 to January 2013: (1) peer-coach and nurse case managed (PC-NCM) program; (2) peer coach (PC) program with brief nurse counseling; and (3) a usual care (UC) program with brief PC and brief nurse counseling. Data were collected on approximately 255 variables on 534 male ex-offenders. The primary outcome of interest was re-arrest within 6- or 12-months post release.

It is common in these kinds of experimental settings for one of a number of regression methods to be used for predictive modeling of outcomes of interest. It is also common for researchers to focus on a particular set of variables in modeling and analysis based on the literature in their fields, which has the potential to leave much of the data unexplored in studies when a large number of variables have been collected. However, modern statistical techniques provide opportunities to include large numbers of candidate predictor variables

from rich data sets. In particular, data mining using machine-learning techniques can be invaluable in identifying characteristics in large data sets that may otherwise be overlooked due to the sheer size or volume of data [66]. This randomized controlled trial can be used to illustrate how machine-learning influenced analytical tools can be used to explore the full richness of the data while also directing attention to a subset of predictors by prioritizing variables in order of importance with regard to predicting a given outcome.

The main purpose of this paper is to compare three machine learning methods based on the top set of variables that are most strongly associated with re-arrest. In this case, choosing the top 20 variables is a convenient choice that illustrates the alternative methods and provides unambiguous basis for comparison across methods. Specifically, we implement the approaches of classification and regression trees (CART), Random Forests (RF) [23] and a newly developed method known as Fuzzy Forests [27]. Fuzzy Forests, an extension of Random Forests, can handle correlated features. Difficulties of fitting and interpreting models in high-dimensional data problems are often exacerbated by the fact that many variables are highly correlated and the correlation structures are not known apriori. In comparing these decision tree methods, there also exist some parametric methods that can handle correlated features. Least absolute shrinkage and selection operator (LASSO) is one such method, which will be explored here as a counterpoint to the decision tree methods. Our primary purpose in this paper is not predictive modeling but rather dimension reduction, starting with the comprehensive list of all variables in the data and reducing it to the top 20 most important variables in relation to the outcome, which then can be used in predictive modeling.

3.3 Background

Data from the randomized controlled trial conducted by Nyamathi et. al [56] was used for this paper, which was funded by the National Institute on Drug Abuse (NIDA), 1 R01

DA27213. The sample included male ex-offenders with a history of drug use prior to their last incarceration and who were between 18-60 years old, living in a residential drug treatment (RDT) program in Los Angeles, and homeless upon discharge from jail/prison. Of the 669 individuals assessed for eligibility and based on age, homeless status, time since release from jail/prison, and drug use prior to recent incarceration, only 600 met the eligibility criteria to be entered into the study. These individuals were randomly assigned to one of the three treatment groups: : (1) peer-coach and nurse case-managed (PC-NCM) program; (2) peer coach (PC) program with brief nurse counseling; and (3) a usual care (UC) program with brief counseling from a peer coach and nurse. Data were collected at 3 different time points: baseline, 6 months and 12 months. However due to missing information on re-arrest, analysis will be based on only 534 male ex-offenders.

3.3.1 Features and Arrest

The baseline questionnaires measure various items such as socio-demographic information, criminal history, drug history, health status, and self-esteem. The various instruments and demographic variables are displayed in Table 1.

Some of the instruments in the study such as the CES-D support the use of composite-score measures that facilitate dimension reduction, with previous studies of reliability and validity suggesting that the composite variables capture the essence of that instrument. In such cases, we considered such composite scores or summary measures to be single variables, representing the entire instrument. For other instruments which could not be reduced to such summary measures, we treated each individual item within the instrument as an individual variable. Variables were excluded from analyses when they had more than 20% missing data or were structurally missing as when an item was not asked of a subgroup due to a questionnaire skip pattern. In this way, there were a total of 210 variables retained in the database.

3.4 Methods

Machine-learning techniques can used to make sense of complex datasets by detecting patterns in the data. There are a variety of machine learning techniques, but this paper will focus on decision trees. Decision-tree methods are non-parametric and non-linear so as to allow for high-order interactions. They also support relatively straightforward interpretation, and their computational efficiency makes decision trees useful not only for predictive modeling but also for variable reduction. All three methods are used to rank variables, in descending order, by their association with re-arrests at 12-months. Since these methods are non-parametric, they can be used in a multitude of situations, with special utility for situations when parametric assumptions do not hold or when the underlying data structure is non-linear or includes high-order interactions.

Alternatives to these decision-tree methods include logistic regression and group LASSO. The results from all five methods are presented, with parametric models yielding findings that can be compared to variable rankings obtained from CART, Random Forests and Fuzzy Forests.

3.4.1 Classification and Regression Trees (CART)

A decision tree is a base learner, that is an approach to classification that provides a foundation for more complex learning strategies. At its heart, a decision tree tries to predict an outcome by recursively partitioning the feature space such that the terminal nodes are homogeneous. That is, it starts with all the data and then examines each possible predictor at each possible split to determine which variable and what cutoff will form the most similar groups in the resulting two daughter nodes (partitions). Within each of the resulting partitions, the process repeats and the variables and various cutoffs are compared to determine which variable best splits each partition into the most similar groups. Each

partition is known as a leaf. The partitions are continually split until the subgroups are sufficiently small or achieve some other user-defined stopping criteria, thus forming a branch of the tree. Inevitably, the first variable chosen at the first split or partition indicates the single most important variable in the dataset. Similarly, the further down the tree the variables are selected, the less important they are in predicting outcomes. As a result of the continued partitioning of the data described by each branch, interactions between the variables on the branch can be readily illustrated. CART is also invariant to monotonic transformations in the variables [67]. As a result, monotonic transformations in the variables do not affect the order and selection of the variables in the tree. Once a partition in a branch can no longer be split, the resulting observations form a terminal node. The terminal nodes are generally classified as the average of the outcome values for all those observations in that node if the outcome is continuous and by a majority vote rule if the outcome is categorical. Sometimes, in the interest of avoiding overfitting, CART procedures include a step of "pruning" a tree based on a criterion to undo a partition.

3.4.2 Random Forests

While CART produces useful graphical representation of the decision tree, it can be unstable, in that even small changes to the data can potentially result in different variables being chosen at a particular split and resulting in a different tree [68]. Ensemble classifiers address this issue by combining multiple trees into the classifier. Random Forests is an example of an ensemble classifier which is presented in the seminal paper by Breiman [23]. Random Forests is able to create multiple trees from the same data by building trees from bootstrap samples with replacement from the data. Each resulting bootstrap sample can be expected to contain 66% of the entire sample. The resulting prediction for each observation is based on the average prediction across the individual bootstrapped trees for a continuous outcome or the majority vote in the case of a categorical outcome. Also as a result of the bootstrap sampling, an out-of-bag (OOB) error rate can be calculated. The OOB error

uses the observations that were not selected in the bootstrap sample as a test set to obtain predictions from the trees that do not contain this observation. Then the error is calculated and averaged across the trees that did not contain that observation.

Since Random Forests utilizes the results of many bootstrapped decision trees, the ease of interpreting relationships between variables is lost. However, Random Forests is able to rank the variables and produce measures of variable importance for each variable. Variable importance can be calculated using either the GINI index or permutation importance. GINI importance is calculated by summing up the values of the GINI index, which serves as a measure of node impurity and which is computed at each split. The GINI index is a measure that in this context can be calculated as the complement of the squared probability of correct classification. Permutation importance determines the importance of a variable based upon the impact of permuting the values of the predictor variable across observations and then averaging the decrease in accuracy across the trees compared to the accuracy of prediction without such permutations. A large change in accuracy implies that the given variable is important since perturbing the observations had a large impact.

Since there was some missingness in the data, missing values were imputed using routines incorporated in the R package for Random Forests. Beginning with medians of continuous variables and the most frequent observed category for categorical variables, values were then iteratively updated utilizing the proximity matrix produced from Random Forests. The proximity matrix indicates how often a set of observations ends up in the same leaf node in a tree. This information is then used as a basis for weighting other observed cases, specifically by calculating the weighted average for continuous variables or by selecting the category from a categorical variable that has the largest average proximity. Five iterations of the Random Forest procedure are used to produce imputed data that can be incorporated in downstream analyses [47].

3.4.3 Fuzzy Forests

While Random Forests creates more stable classifiers than its single tree counterpart, the resulting variable importance measures can be biased in the direction of attaching too much importance to variables correlated with an important predictor ([75], [53], [52], [15], [38]). To counter this, Fuzzy Forests are designed to provide relatively unbiased rankings of variable importance in the presence of highly correlated variables. In order to achieve this, Fuzzy Forests first separates variables into groups or modules that have similar pairwise correlations using a weighted correlation network. Using these modules as starting points, the next stage of Fuzzy Forests uses recursive feature elimination Random Forests, a procedure that is iteratively performed in each group so as to discard the least predictive set of variables until the best set, according to the OOB error rate, from that group remains. At the final stage, the most predictive variables from each module are pooled together and another Recursive Feature Elimination Random Forest is run to produce a final set of variables that are predictive of the outcome.

Fuzzy Forests was developed primarily in the context of flow cytometry, genetics, proteomics, and other bioinformatics data settings where machine learning methods can be used to achieve dimension reduction. Typically, all the variables in contention are continuous. In the context of social experimental settings, given the large number of relevant variables, dimension reduction remains an important consideration in predictive modeling. However, variables are typically a mix of continuous and categorical. While measures of association can still be computed and modules can be created to run Fuzzy Forests as described above, there usually exists a natural grouping among variables which can be utilized in the creation of modules. Experimental settings are generally guided by clusters of constructs that are related. So, it seems natural to consider such clusters as modules in Fuzzy Forests.

For the present data, to group the 210 candidate predictor variables, eight modules were created with the following labels: demographic characteristics, knowledge/attitudes of

sexual diseases, support during childhood and from family, psychological state, health, drug use / sexual activity, criminal history, other. Variables that did not fit together into any meaningful category were put in the other category; examples include if the individual was in jail or prison and how often one prays.

3.4.4 Module-wise two-step Logistic Regression

With 534 observations and 210 candidate predictor variables, complete and quasi-complete separation are concerns that can be expected to arise in a logistic regression analysis. To further illustrate the necessity and benefits of accounting for the correlation structure or to introduce groups of similar variables, lessons learned from Fuzzy Forests will be used to avert complete separation in the data. First, a stepwise logistic regression is performed on each group of predictors within a module according to the Fuzzy Forests framework. Then the retained significant variables from the first step are combined into a final forward-selection logistic-regression model.

3.4.5 Group LASSO

Similar to the logistic regression approach, group LASSO [85] penalizes the coefficients such that the coefficients corresponding to insignificant variables are shrunk to 0. LASSO requires that all levels in a categorical variable are dummy coded producing indicator variables for each level, excluding the reference category. Group LASSO ensures that all indicator variables pertaining to a specific variable either all have zero or non-zero coefficients. To avoid ambiguity surrounding small-absolute-value coefficients, the LASSO procedure was implemented in a way that retained only those coefficients larger than 0.1 in absolute value, discarding the remaining variables. The retained variables were then included in a logistic regression to obtain corresponding p-values.

3.4.6 AUC

The above methods can all be utilized for dimension reduction. To compare these three machine learning methods, an objective measure such as the area under the curve (AUC) of a receiver-operator characteristic (ROC) curve can be used. AUC can be interpreted as the probability of a correct ranking ([20], [41]). As an additional comparison, a hold-out sample comprising 10% of the original data was selected and the hold-out sample was then used to determine the accuracy of the predicted number of re-arrests based on comparing model predictions to the true number of re-arrests.

3.5 Results

Descriptive details of the sample can be found in the report on re-arrest of recently released homeless offenders by Nyamathi et. al [56]. Here, we focus on the results comparing CART, Random Forests, Fuzzy Forests, Module-wise two-step logistic regression, and Group LASSO.

Initial exploration of the data revealed that time spent in a residential drug treatment center is such a strong predictor of re-arrest relative to the other variables, in that it was the top ranked predictor across all the methods (correlation =0.48). So when removed, it allows other variables to have the chance to compete for inclusion into a tree and thus improve the variable importance of other variables by potentially allowing unexpected and more interesting variables to be highlighted in this setting. The resulting models presented below all exclude the time in residential drug treatment programs from the analysis.

The results of the three decision tree methods are presented in Figures 1-3. Fuzzy Forests determines that the best predictors of re-arrest are how willing the individual is to watch an execution and also how strong of an urge the person has to help when they see

someone in distress. The remaining top twenty variables have relatively lower variable importance scores. The CART results do not seem to cluster together, but the most important variable corresponds to how strong of an urge they have to help when they see someone in distress. Less important variables correspond to number of cigarettes per week, if the individual is a current smoker, how willing one is to watch an execution, if the individual very much enjoys and feels uplifted by happy endings, contract type measuring participants most recent experience with the California Department of Corrections and Rehabilitation's treatment, and also whether the individual and their spouse, significant other, or partner helped each other with problems in the 6 months prior to incarceration. The variable importance values for Random Forests show a similar clustering to Fuzzy Forests with the best predictors similarly being identified as how willing an individual is to watch an execution and how strong of an urge they have to help when they see someone in distress. Among these machine learning methods, how willing they are to watch an execution and how strong of an urge they have to help when they see someone in distress are common highly ranked variables, thus reinforcing their importance regardless of which machine learning method is chosen.

Table 2 also presents the results of Fuzzy Forests, CART and Random Forests, but only includes the set of 15 stable variables from each of these methods, where stable variables are determined by taking 5 different seeds and selecting the variables consistently chosen from each method. We then took the 15 stable variables and conducted a logistic regression for each model. Many of the variables were not statistically significant. Each of these three machine-learning methods identified four common significant variables; not watching an execution, number of times went to juvenile hall, whether or not they are a current smoker, and the degree that helpless old people have an emotional effect on them were all significant predictors of re-arrest at the 0.05 level. CART additionally identified whether or not their spouse / significant other / partner helped each other with problems in the 6 months prior to incarceration and how much time they were in prison for the current incarceration. Along with the common four significant variables, Fuzzy Forests also identifying contract type.

The module-wise two-step logistic regression and group LASSO models are presented in Table 3. All but six of the top 20 selected variables from module-wise two-step logistic regression were significant predictors of re-arrest at the 0.05 level. Group LASSO only had five variables with non-zero after implementing a cutoff value of 0.1 for coefficients. Group LASSO results show that the degree that helpless old people have an emotional effect on them, strong urge to help when someone is in distress, not wanting to watch and execution and contract type are significant predictors of re-arrest.

Overall comparisons of the models can be found in Table 4 and 5. Table 4 includes the OOB error rates for Random Forests and Fuzzy Forests along with the AUCs from the five methods. The top 15 variables from each of the machine-learning method were incorporated into separate logistic regression models to obtain an AUC for each method, and AUC values were similarly obtained for a module-wise two-step logistic regression and a group LASSO procedure (which yielded five predictor variables). In Table 4 the OOB error rate is slightly higher in Random Forests as compared to Fuzzy Forests. Among the machine learning methods, Fuzzy Forests slightly outperforms Random Forests in terms of both AUC and OOB error rates. Among the machine-learning methods with a module-wise two-step logistic regression based top 15 variables in the models and the group LASSO procedure which yielded five predictor variables, the largest AUC was obtained from the module-wise two-step logistic regression model, with an AUC = 0.81, followed by CART and then Fuzzy Forests with an AUC of 0.78.

However, when a 10% hold-out sample is used as a test-set, the misclassification is lowest for Fuzzy Forests, indicating that while initially the module-wise two-step logistic regression seems to slightly outperform all the other models, when it comes to using the model to predict on a different dataset, it incorrectly predicts the outcome at a higher rate than Fuzzy Forests. These findings suggest that the two-step model may be over-fitting the data. In the test set, group LASSO performs better than the module-wise two-step logistic regression with a lower misclassification rate, but has a comparable AUC. Random Forests

and CART result in the highest misclassification rates and the lowest AUCs out of the five methods.

3.6 Conclusion

In this paper, we compare Fuzzy Forests, a novel machine learning algorithm for ranking the importance of variables in high-dimensional classification and regression problems, to the popular Random Forests, CART, a modified version of logistic regression and group LASSO.

In our data, we found that the machine-learning methods Random Forests, Fuzzy Forests, CART are comparable in detecting the important variables in the data. However, these methods are able to detect more interesting variables once the variable with the highest signal is removed. In this case, residential drug treatment is so strong at partitioning the data into similar outcomes that the other variables were not adding much information. By removing this strong variable, other variables can be allowed to compete and may be included in more trees, thus improving their variable importance.

While the module-wise two-step logistic regression does seem to perform well in predicting outcomes and produces high AUCs, the predictors were only identified after building upon the theory utilized in Fuzzy Forests. This allowed for a large percentage of significant variables to be identified among the top twenty predictors. If existing methods such as LASSO are used, it is only able to identify a limited set of predictors, at best, upon which to focus future research. In situations where similar data could be gathered, methods such as logistic regression and group LASSO would not necessarily be anticipated to provide new insight into relationships of interest between predictors and outcomes. In the present context, machine-learning methods, and Fuzzy Forests in particular, outperformed other decision tree methods along with logistic regression and LASSO.

Overall, we conclude that these machine learning techniques are comparable in dimension reduction, with the new Fuzzy Forests technique being preferred over the other methods in its capacity to deal with correlated variables and its performance in the test sample. The selected variables from these machine-learning techniques can then be used to construct a predictive model for the outcome. The benefit of these machine-learning techniques is that all variables in a database are assessed in relation to the outcome with a ranking of their relative importance for prediction. In the absence of such methods, if we were to go directly to predictive modeling, then we would encounter challenges in dealing with a large number of variables. Model-selection is topic of intense interest in many fields. Indeed, considering the substantial resources that go towards conducting these studies, we are obligated to leverage these big data and conduct comprehensive analyses of our data. Machine learning techniques such as those discussed in this paper enable researchers to do just that.

Table 3.1: Baseline Measurements and Questionnaires

Questionnaire (number of items within instrument)	Measurements
Demographics (19)	Examples: Race, Ethnicity, Children, Education, Employment
HIV/AIDS knowledge (HIVKNOW) (17)	Knowledge and attitudes about HIV/AIDS
Knowledge and Attitudes about Hepatitis B (HBVKNOW) (3)	Knowledge and Attitudes about Hepatitis B
Childhood and Family Background (31)	Examples: Living alone, Interactions with spouse / significant other or domestic partner
Peer Relations (24)	Relationship Examples: Gang member, Types of support
Mental Health Index (5)	Likert-scaled items on how an individual has been feeling in the past month
Center for Epidemiologic Studies Depression Scale (CES-D) (10)	Depression Scale based on aggregating 10 Likert-scaled item responses
Brief Cope (16)	General methods of coping with stress
Self-Esteem Inventory (23)	How one feels about oneself
Balanced Emotional Empathy Scale (BEES) (7)	Describe how one would act in certain situations
Brief Symptom Inventory Hostility Subscale (BSI) (5)	Describe how much a problem bothered them
Health and psychological status(29)	Examples: Overall health, effects of drugs and drug treatments

Table 3.1 continued from previous page

Drug and treatment History(29)	Drug and treatment history
CAGE screener and augmented screener items(8)	Drinking questions
Sex Behavior(3)	Sexual behavior
Criminal History(51)	Criminal History

Table 3.2: Demographic Table

Demographic Variable	Mean(SE)
Age	39.9(0.45)
	N(%)
Education	
$\leq$ High School	479 (89.7%)
At least some college	55 (10.3%)
Marital Status	
Married/living together	58 (10.9%)
Never married/separated/ divorced/ widowed	476 (89.1%)
Re-arrested	
Yes	331 (62.0%)
No	203 (38.0%)

Table 3.3: Comparison of Fuzzy Forests, Random Forests and CART

Fuzzy Forests	CART	Random Forests
Not watch execution***	Strong urge to help when I see someone in distress	Not watch execution**
Strong urge to help when I see someone in distress	Number of cigarettes per week	Strong urge to help when I see someone in distress
Number of times at juvenile hall*	Current smoker*	I very much enjoy and feel uplifted by happy endings
I very much enjoy and feel uplifted by happy endings	Not watch execution***	Number of cigarettes per week
Number of cigarettes per week	I very much enjoy and feel uplifted by happy endings	Number of times at juvenile hall*
I hardly ever cry when watching a very sad movie	Contract Type	Current smoker*
Current smoker*	You and your spouse /significant other /partner helped each other with problems in the 6 months prior to incarceration*	I hardly ever cry when watching a very sad movie

Table 3.3 continued from previous page

Time spent in juvenile hall	Average mental health score	Contract type
Number of arrests before 18 yrs old	Total Self Esteem Score	Time spent in juvenile hall
Contract type*	Helpless old people do not have much of an emotional effect on me**	Number of arrests before 18 yrs old
Number of times used marijuana in the 6 months before incarceration	General knowledge of HIV	Helpless old people do not have much of an emotional effect on me
Age	How much time have you spent in jail	Age
Age at first arrest	How much time have you been in prison for your current incarceration*	Number of times used marijuana in the 6 months before incarceration
Helpless old people do not have much of an emotional effect on me**	Number of times at juvenile hall*	6 months prior to arrest, how many times did you commit probation/parole violations**
Number of times at residential drug program, excluding alcohol treatment	Age	The sadness of a close one easily rubs off on me

Table 3.3 continued from previous page

Note: Contract type measuring participants most recent experience with the California Department of Corrections and Rehabilitation(CDCR) treatment programs while in prison

* <.05, **<.01, ***<.001 significance level from logistic regression

Table 3.4: Results of Module-wise two-step logistic regression and Group LASSO

Module-wise two-step Logistic Regression		Group LASSO	
Variables	Chisquare P-value	Variables	Chisquare P-values
I very much enjoy and feel uplifted by happy endings	0.541	I very much enjoy and feel uplifted by happy endings	0.1
Not watch execution	<0.001	Helpless old people do not have much of an emotional affect on me	0.007
Current smoker	<0.001	Strong urge to help when I see someone in distress	0.014
Helpless old people do not have much of an emotional effect on me	<0.001	Not watch execution	<0.001
Total number of times arrested	0.038	Contract type	0.037
Ever used cocaine	<0.001		
Ever went to a residental drug/program excluding alcohol treatment	0.016		

Table 3.4 continued from previous page

You and your spouse/significant other/partner helped each other with problems in the 6 months prior to incarceration	0.003
Contract type	0.005
In the last 6 months prior to incarceration, close relationships are between family or friends	0.008
Number of times used marijuana in the 6 months before incarceration	0.011
Number of children	0.011
Times committed shoplifting/larceny/ embezzlement in the 6 months before arrest	0.019
Frequency used heroin/meth in the last 6 months before this incarceration	0.012
Time spent in juvenile hall	0.085

Table 3.4 continued from previous page

You need help getting a license to stay out of prison	0.046
Times committed robbery/attempted robbery/mugging in the 6 months before arrest	0.076
Number of close friends and relatives you currently have outside of prison	0.087
How much time have you been in prison for your current incarceration	0.134
Ever used meth/heroin/cocaine	0.139

Table 3.5: Comparison of Models using Fuzzy Forests, Random Forests, CART, Module-wise two-step Logistic Regression and group LASSO

	Fuzzy Forests	CART	Random Forests	Module-wise two-step Logistic Regression	Group LASSO
OOB	29.59	–	32.58	–	–
AUC from top / stable 15 variables	0.779	0.780	0.775	0.814	0.749

Table 3.6: Comparison of Models using a 25% hold out sample

	Fuzzy Forests	CART	Random Forests	Module-wise two-step Logistic Regression	Group LASSO
Misclassification	0.264	0.396	0.377	0.302	0.283
AUC	0.716	0.584	0.580	0.693	0.694

Figure 3.1: CART

Figure 3.2: Random Forests

Figure 3.3: Fuzzy Forests

CHAPTER 4

Genetic variants associated with Lithium Response in bipolar Disorder

4.1 Abstract

4.1.1 Background:

Bipolar disorder has a large global health burden. Lithium treatment is an effective treatment for some patients experiencing bipolar disorder. However, previous research has revealed that approximately 30% of patients do not respond to lithium. Anecdotally, it has been suggested that patients whose family members who responded to lithium treatment are more likely also to be lithium responders, suggesting a possible genetic component.

4.1.2 Data:

Two datasets were collected from the GAINS network of genetic databases. All subjects were on lithium treatment, and their lithium response was recorded as determined using the Alda Scale. The larger database, a retrospective dataset, was used as a training set, and the smaller database, the prospective dataset, was used as a test set. Each dataset contains recorded information on 248 SNPs believed to be enhanced for lithium response as

well as demographic information on gender and self-reported race-ethnic identity and clinical information including family history of bipolar disorder and lithium response.

4.1.3 Methods:

We used the Fuzzy Forests algorithm, a recently developed machine-learning approach to examine predictors of lithium response. We trained the data on the retrospective data and tested the model using the prospective dataset. We also performed univariate logistic regressions incorporating each of the potential predictors.

4.1.4 Results:

Among the 248 SNPs and demographic characteristics explored between the retrospective and prospective datasets, only rs2241382 was selected in the top twenty best predictors for each dataset. Other predictors that were important, but were not chosen among the retrospective and prospective datasets, were gender and family history of bipolar disorder. In univariate logistic regression, history of bipolar disorder and 10 SNPs yielded unadjusted p-values less than 0.05, with the most significant having an unadjusted p-value of 0.003, but none of these predictors remained significant after multiple-comparison adjustment.

4.1.5 Conclusion:

We were unable to detect a genetic signal for lithium response that would reliably be reproduced in other settings. Machine learning cannot overcome defects in the study design. Our study, which was limited by power to detect low effect size, cannot rule out smaller genetic effects.

4.2 Introduction

According to the National Institute of Mental Health, bipolar disorder is defined as a brain disorder that causes unusual shifts in mood, energy, activity levels, and the ability to carry out day-to-day tasks [9]. It is characterized by mood episodes of mania and/or depression. Some symptoms of bipolar disorder in the manic state include increased activity levels with increased energy, trouble sleeping, being agitated, feeling high, and talking fast about many topics. Similarly in the depressive state, symptoms include feeling very sad, with little energy, having trouble sleeping, feeling worried or trouble concentrating, change in eating habits, disinterest in activities, along with suicidal ideation. However, the presence of the disease is not limited to only situations when these extreme symptoms are present. Bipolar disorder can also present as less extreme, such as with a less severe form of mania called hypomania. With hypomania, a person may feel fine and function well, but others may realize there are changes in mood or activity level. If left untreated, hypomania may lead to the more severe forms of mania or depression. When the mood episode includes both mania and depression, also known as an episode with mixed features, one may feel very sad and simultaneously energetic. On occasion, in a severe episode, a person may experience some psychotic symptoms that include hallucinations, which can lead to being misdiagnosed as schizophrenia. In general, episodes, either manic or depressive, recur over time. Between episodes, many people do not show any mood changes, but others have lingering symptoms [9]. The severity of symptoms varies by person, but the average onset of bipolar is at age 25 [7]. Each year, 2.6% of U.S. adults are diagnosed as bipolar with 82.9% classified as severely bipolar [10].

NIH also classifies the diagnosis of bipolar disorder into bipolar I, bipolar II, cyclothymic disorder, and other specified and unspecified bipolar and related disorders. Bipolar 1 is defined by a manic episode that is sustained for at least 7 days or is so severe that it requires an immediate visit to the hospital. Depressive episodes can occur and will usually be sustained for at least two weeks. An episode having both depressive and manic symptoms can

also occur. Bipolar II is characterized by depressive and the less severe, hypomanic episodes. Cyclothymic disorder includes periods of hypomanic symptoms and periods of depressive symptoms lasting for at least 2 years or 1 year in children and teens. Other specified and unspecified bipolar related disorders are classified as having bipolar symptoms but do not meet the classifications of the previous categories [9] Bipolar classification also includes rapid cycling which is defined as having at least four episodes of mania and/or major depression each year [18].

Lithium is a mood stabilizer and is a primary treatment for bipolar disorder. While many patients respond to lithium, approximately 30% of patients are non-responders or partial responders [34]. While bipolar disorder tends to be common among family members, some studies have suggested there may be a genetic component to bipolar disorder. Evidence for a particular gene or set of genes has been inconclusive ([30], [29]); however, there have been many studies identifying a few susceptibility loci such as 12q23q24 that are associated with bipolar disorder ([30], [35]). There are also some studies suggesting that lithium response might also be heritable, with patients having a family history of bipolar disorder more likely to respond to lithium treatment [46].

Evidence of a genetic link between bipolar disorder and lithium response could provide valuable insights into the mechanisms of the lithium response and yield insights into treatment as well as into possible new treatments for those that do not respond. There have been several candidate gene studies on the genetics of lithium response that so far have not yielded reproducible results ([50], [25]).

Genome-wide association (GWA) studies provide information on the entire genome and allow researchers to determine which genes are associated with the bipolar disease without relying on prior knowledge [26]. The Foundation for the National Institutes of Health, Inc. (FNIH) has a public-private partnership with the Genetic Association Information Network (GAIN). GAIN provides support for GWA studies that identify DNA variations associated with a specific common disease [11]. GAIN funded the GWA study of bipolar

disorder, which was conducted by the Bipolar Genome Study (BiGS) in 2006. BiGS was part of the larger Bipolar Disorder Consortium that collected data on over 3,500 subjects and was funded by the National Institutes of Mental Health (NIMH) [69].

The GAIN dataset includes a subset of 1001 cases of bipolar disorder and 1033 controls of European Ancestry genotyped through the GAIN initiative [55]. Additionally, samples collected using similar mechanisms were used to obtain another subset called the Translational Genomics (TGEN) sample. The TGEN sample contains 1190 genotyped bipolar disorder cases from the Bipolar Genome Study along with 401 controls. The sample collection from the GWA study included genotyping using the affymetrix genome-wide human SNP array 6.0 [37].

Among the genetic data already mentioned, additional information was collected using a drug response questionnaire. This questionnaire included information on whether the subjects had taken lithium and whether they responded to lithium using the Alda scale. The subset of the data analyzed includes 326 retrospective and 97 prospective individuals and 248 SNPs for which there is data on lithium response from these two different datasets. Limited demographic information in the data includes family history of being bipolar, ethnicity, and gender. The prospective data was collected to test the reproducibility of the genes found in the retrospective study.

In addition to being used to treat bipolar disorder, lithium can also be used as a prophylactic [81]. While lithium can be effective, not all bipolar patients respond well to lithium treatment. The lithium response among patients is clinically different in responders versus non-responders. Lithium responders typically exhibit a pattern of non-rapid cycling [79]. Individuals with a family history of first-degree relatives whose bipolar disorder responds to lithium are themselves more likely also to respond to lithium [39]. This implies that there might exist heritability of the response to lithium treatment.

Thus, building on this premise, the current investigation aims to analyze genetic data

to identify potential SNPs related to lithium response in the retrospective data set and validate it on the prospective dataset. However, because there may be associations among the SNPs and because there are a large number of SNPs relative to the number of subjects, we use the Fuzzy Forests algorithm [28] to determine predictors of lithium response and determine if any other susceptibility genes can be identified.

4.3 Methods

The GAIN databases were collected from 10 sites: Indiana University (with satellite sites at University of Louisville and at Wayne State University in Detroit), Johns Hopkins University, the NIMH Intramural Research Program, Rush-Presbyterian Medical Center in Chicago, University of California at Irvine, University of California at San Diego, University of Chicago, University of Iowa, University of Pennsylvania, and Washington University in St. Louis. Specific, uniform ascertainment rules were used at all sites. Subjects were identified by screening admissions at local treatment facilities or by advertisement through advocacy groups, Web sites, and professional organizations. Participants were asked to give informed consent for interview, a blood specimen for DNA and cell lines, and permission to contact relatives; the study was approved by institutional review boards at each participating site.

The genome scan was performed at the Center for Inherited Disease Research by use of automated fluorescent microsatellite analysis. PCR products were sized on an ABI 3700 Sequencer. The marker set used was a modification of the Cooperative Human Linkage Center version 9 marker set (391 markers, average spacing 9 cM, average heterozygosity 0.76). The error rate, based on 17,707 paired genotypes, was 0.05%. The overall missing data rate for the 471,032 total genotypes was 3.75% per genotype. All genotyping was performed blind to clinical status.

We analyze two separate subsets of data collected from the GAIN database network

where the Alda scale was available. All patients in the study had a DSM-III or DSM-IV diagnosis of bipolar disorder. Data were gathered on gender, ethnic origin, family history and a total score on the Alda scale. Patients were on lithium for at least 6 months before the assessment on lithium response was done. The Alda scale quantifies symptom improvement. Patients who scored a 7 and above on the Alda scale were coded as lithium responders and below 7 as non-responders for both datasets. The first database, which we call retrospective, was used as our training dataset and the second dataset, which we call prospective, is our validation dataset. Both datasets were collected from the GAIN database.

Genetic data for each locus is given in terms of allele frequencies indicate the frequencies that indicate the number of copies of specific genes at each locus. The frequency ranges from 0 to 2 where 0 indicates that there is no mutation from wild-type at either allele. A score of 2 indicates that both alleles include a mutation. Table 1 lists the demographic features of both the retrospective and prospective datasets, as well as the p-values for the difference between the datasets. P-values were calculated using Fishers Exact Test. Note that the datasets did not differ on the percentage of lithium responders. However these two datasets differ in their proportion of males as well as self-reported ethnicity/race.

The retrospective dataset was trained using Fuzzy Forests. Fuzzy Forests is an extension of the machine learning algorithm Random Forest [23] that can give less biased variable importance scores when there is high correlation among the predictors. Briefly, Fuzzy Forests performs a weighted correlation network to form modules or clusters of the input data. Within each module, Recursive Feature Elimination Random Forests is performed to obtain the set of SNPs that are most associated with lithium response. This set of SNPs is obtained by iteratively removing the SNPs with the lowest variable importance. Variable importance is measured by permuting the observations of a variable in the out-of-bag (OOB) dataset, which refers to cases not included in the given bootstrap sample. Once permuted, the data are then passed down the tree. The variable importance is the mean decrease in the number of correctly predicted classes using the original OOB sample and the permuted

sample.

This is done until a preset limit on the number of parameters per module is reached. After the best set of SNPs are selected from each module, they are pooled together for a final recursive feature elimination random forest to obtain the final set of best SNPs[[28], [27]].

The retrospective allele information is used to train the Fuzzy Forests algorithm from which the prospective allele information was used to determine the accuracy of the model obtained from Fuzzy Forests. For comparison of ranks of variable importance among the SNPs, Fuzzy Forests was performed on both the retrospective and the prospective datasets to identify the commonalities between both in terms of the important SNPs that were identified. Note that this is not a contamination of the analysis, as the result Fuzzy Forests object was not used in the evaluation of the predictive performance. The Fuzzy Forests was specifically done to look for the commonalities in genes selected by the algorithm post-hoc to examine common patterns of the genes in both datasets.

After the algorithm is trained on only the retrospective data, the misclassification rate in the prospective data will be used to summarize the accuracy of the algorithm. To handle any missing observations, the retrospective data is first imputed using five iterations of random forest imputation prior to the utilization of Fuzzy Forests. We also performed more traditional logistic regression to test the association of allele frequencies and lithium response.

4.4 Results

Table 1 lists the demographic information for both the retrospective and prospective samples. Note that the lithium response rate is similar in both datasets. However, notice

that the retrospective and the prospective databases differ significantly in terms of their gender and race/ethnicity. The retrospective data has a relatively balanced gender ratio with 52.1% male and 47.9% female. However the prospective data, which is smaller, has significantly more males, comprising almost 88% of the sample ($P < 0.0001$ by Fishers Exact Test). Similarly the prospective dataset is more racially diverse. In the retrospective dataset, African Americans comprise 2.8% of the dataset and those of European descent comprise 87% compared to 10.3% and 70.1% in the prospective dataset. These differences in ethnic/racial distribution of subjects are significant at p=0.0007 by Fishers Exact Test.

We ran Fuzzy Forests on the retrospective (training) data set and examined the top SNPs in predicting lithium response in bipolar disorder. Potential predictors were family history, gender, and SNPs from 23 chromosomes. We found SNP rs2194368 had the highest variable importance followed by rs1992248, rs1410113, rs4949915 and rs162159 as the respective top five important SNPs. The Variable Importance Plot is given as Figure 1 and lists the top 20 SNPs. The results of the Fuzzy Forests algorithm resulted in an OOB error rate of 30.7% using the retrospective data. The results were similar when adjusting the tuning parameters.

We then tested the prospective data using the Fuzzy Forests object trained on the retrospective data. There was a 55.67% misclassification rate of the predicted lithium response rates when we tested the model using the prospective data. We then ran a Fuzzy Forests model on the prospective data to ascertain the top SNPs in predicting lithium response in this data set. In the prospective data, the SNP rs903085 was the most important variable among the prospective data, followed by rs295886, rs12538005, family history of bipolar, and rs11160586. Figure 2 plots the top 20 important SNPs in predicting lithium response in the prospective data. The OOB error rate using this data set was 25.77%.

Because some demographic characteristics were significantly different between datasets, we examined the misclassification rates of the training dataset after subsetting by gender and testing on the prospective data set. When we subset the data using only males,

the misclassification rate was approximately 65%. When looking at only females, the misclassification rate was approximately 42%. We did not have the sample size to subset on race/ethnicity other than those of self-reported European ancestry. When subsetting on this category, the misclassifcation rate was approximately 62%. We found similar high misclassification rates when we subset on family history of bipolar. Table 2 lists the misclassification rates for each of these sub-analyses.

There was only a single overlapping SNP out of the top twenty selected (rs2241382) between the retrospective and prospective datasets. It is interesting to note that Family History was selected as being important in the prospective dataset, but not amongst the retrospective data.

We then performed univariate logistic regressions using the binary phenotype of lithium response as the outcome of interest. In the retrospective dataset, neither model resulted in significant results using the Bonferroni multiple comparison significance threshold of $p < 0.000098$. Comparatively, among the twenty univariate logistic variables in the retrospective data with the smallest p-values, there are 12 in common with the results from Fuzzy Forests with 8 of these variables being in the top ten ranked in Fuzzy Forests. For the prospective data, there were 14 common variables amongst the top twenty selected from Fuzzy Forests, with the top ten ranked variables also having among the lowest p-values from the univariate logistic results. Table 3 lists the results of 20 univariate logistic regression from each dataset with the smallest p-values. The full set of univariate logistic regressions from each dataset is located in the Appendix, Table A1. In univariate logistic regression, history of bipolar disorder and 10 SNPs yielded unadjusted p-values less than 0.05, with the most significant having an unadjusted p-value of 0.003, but none of these predictors remained significant after multiple-comparison adjustment.

4.5 Discussion

Despite none of the univariate logistic regression results being significant after multiple-comparison adjustment, the overlap in ranking of the p-values with the ranks obtained from Fuzzy Forests suggest that univariate logistic regression and Fuzzy Forests are both finding the same signal in each dataset. However since the top-ranked variables are ranking different sets of predictors for each dataset, we are only able to predict but not generalize the findings of the training set onto the test set. It is possible that systematic differences between the retrospective and prospective datasets influenced the results.

Despite exploring multiple methods, we were not able to convincingly reproduce findings from the retrospective (training) data set using the testing (prospective) dataset. One statistical issue is that the sample size was small compared to other genetic studies. Another statistical issue is that significant differences in the testing and training datasets in terms of gender and racial/ethnic background suggest the possibility of population stratification which can be a confounder in genetic studies and might have played a role in our lack of significant findings. Comparisons of misclassification rates across different strata indicate that the predictive accuracy of Fuzzy Forests is not greatly improved by restricting attention to subsets of the data, reflecting a lack of evidence for consistent signals in the data.

Genetic architecture is complex, and finding specific polymorphisms associated with bipolar disorder or lithium response has been challenging. Previous genome wide association studies have examined associations of millions of common SNPs. Such studies have consistently found that individual SNPs exert small effects on genetically complex traits [[83], [77], [62]]. For example, a recent study of depression using GWA studies found a strong signal in SNP rs12552 with an odds ratio of $1.044 (P = 6.07x10^{-}19)$. To reliably replicate this findings would require a sample size of 34,100 individuals to be able to detect this signal with 80% power at an alpha level of 0.05, assuming a balanced case-control design [83]. In our analysis, even if we combine the datasets, we have fewer than 400 subjects, suggesting

that we are underpowered to detect even large genetic effects. Machine learning, for all of its sophistication, cannot get around limitations in statistical power when investigating small effect sizes.

Further confounding the results, lithium response may be a physiologic phenomenon instead of a primarily genetic phenomenon. Tobe et al. (2019) mapped the lithium-response pathway, which controls the phosphorylation of CRMP2, and although toggling between inactive (phosphorylated) and active (non-phosphorylated) CRMP2 is physiologic, the set-point in lithium response bipolar disorder is abnormal. Lithium (and other pathway-modulators) normalize that set-point. Hence, bipolar disorder might not be governed by a gene but by the posttranslational regulation of a developmentally critical molecule [78]. Given this finding, it is not surprising that we were not able to find a distinct genetic signal with only SNP data.

4.6 Conclusion

Bipolar disorder affects approximately 5 million adults in the US. Bipolar disorder is a set of complex mood states that varies greatly between some patients, with some people experiencing rapid cycling while others spend more time in either depressive or manic states. The diagnosis of bipolar often takes time and many patients can be misdiagnosed before their final diagnosis is made. Lithium is usually the primary treatment after bipolar is diagnosed but has significant limitations. There can be significant side effects including nausea, muscle tremors, weight gain, and birth defects and many patients stop taking their medication due to this. Roughly only 1/3 of patients with bipolar disorder respond to lithium and usually the effect is only found through a very lengthy trial and error process. So knowing apriori which patients would respond to lithium, would reduce some of the morbidity associated with this disorder. This is why we undertook the study to see if we can find a genetic basis to lithium response.

Lithium response to bipolar disorder appears to also be complex and most likely encompasses many different mechanism and pathways not to mention environmental exposures. We utilized a novel machine learning algorithm to determine if there exists a genetic signal in lithium response to bipolar disorder. We examined SNP data in both a retrospective and prospective dataset, but we were unable to reproduce findings from one dataset in the other. The inability to replicate the findings could be due to several reasons: 1) Given the recent findings that the regulation of gene products is implicated in bipolar disorder, perhaps there is not a strong enough genetic signal in SNPs to be able to accurately predict lithium response in a more generalizable population. 2) Our sample size is small for a genetic study. 3) Known differences in the gender and ethnicity profiles between the prospective and retrospective datasets could lead to artifacts in the genetic signal. We found overlap in SNPs in either cohort only when we used an extremely liberal cutoff, suggesting that such findings might not be reproducible. Future work should include larger genetic studies that include data on gene phosphorylation and methylation. Machine learning, for all of its sophistication and promise, cannot overcome limitations associated with statistical power when investigating small effect sizes. Future genetic studies exploring possible contributing factors to bipolar disorder should anticipate modest effect sizes and should not rely on machine-learning techniques without due consideration of statistical power.

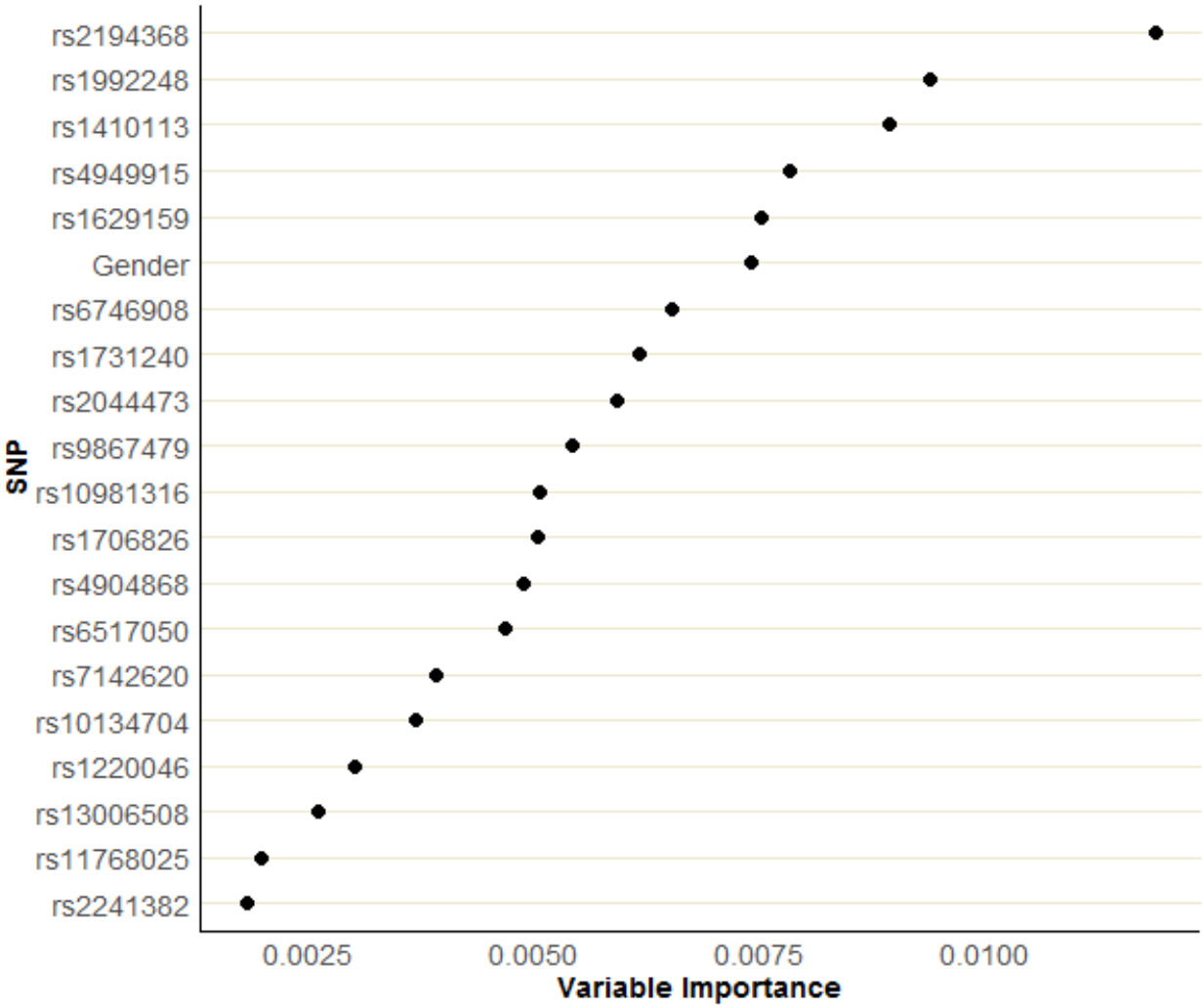

Figure 4.1: Variable Importance measures of SNP using the Retrospective Dataset. SNPs are displayed by rank with the most important SNP in the top position.

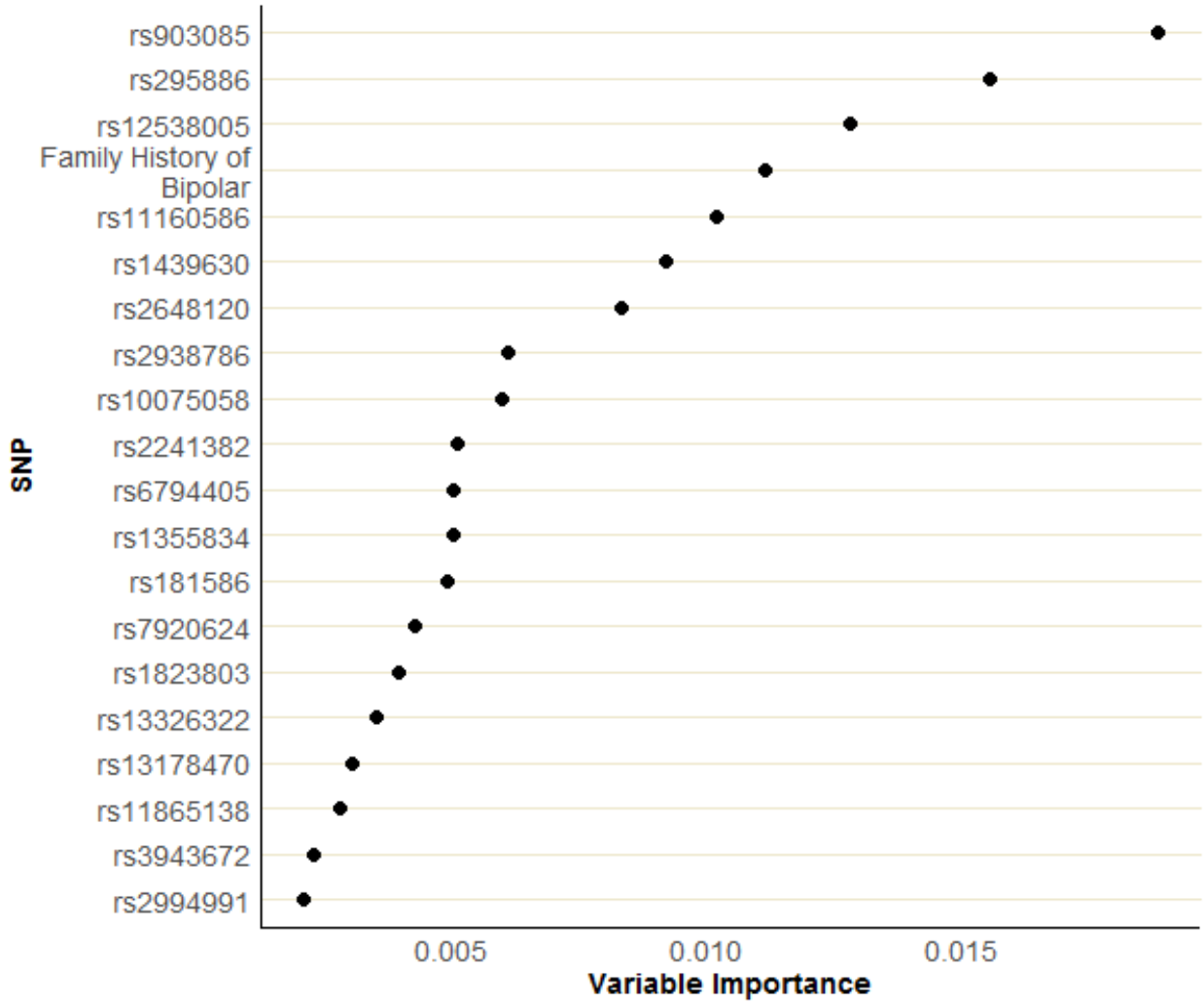

Figure 4.2: Variable Importance measures of SNP using the Prospective Dataset. SNPs are displayed by rank with the most important SNP in the top position.

Table 4.1: Demographic Variables for each cohort by Lithium Response. P values were calculated by Fishers Exact Test

	Retrospective N=326				Prospective N=97				Fisher's P-value
	Overall	Lithium Response		Fisher's P-value	Overall	Lithium Response		Fisher's P-value	
		Yes	No			Yes	No		
Lithium Response		158 (48.5)	168 (51.5)			56 (57.7)	41 (42.3)		0.13
Family History of Bipolar				0.01				0.04	1
Yes	160 (49.1)	66 (41.8)	94 (56)		47 (48.5)	22 (39.3)	25 (61)		
No	166 (50.9)	92 (58.2)	74 (44)		50 (51.5)	34 (60.7)	16 (39)		
Ethnicity				0.44				0.7	0.0007
AA (African American)	9 (2.8)	5 (3.2)	4 (2.4)		10 (10.3)	4 (7.1)	6 (14.6)		
EA (European Ancestry)	285 (87.4)	134 (84.8)	151 (89.9)		68 (70.1)	40 (71.4)	28 (68.3)		
Easia (East Asian)	3 (0.9)	3 (1.9)	0 (0)		2 (2.1)	2 (3.6)	0 (0)		
HNA (Hispanic)	13 (4)	7 (4.4)	6 (3.6)		10 (10.3)	6 (10.7)	4 (9.8)		

Table 4.1 continued from previous page

O	16	9	7		7	4	3	
(Other)	(4.9)	(5.7)	(4.2)		(7.2)	(7.1)	(7.3)	
Gender				0.03				1 <0.0001
Male	170	72	98		85	49	36	
	(52.1)	(45.6)	(58.3)		(87.6)	(87.5)	(87.8)	
Female	156	86	70		12	7	5	
	(47.9)	(54.4)	(41.7)		(12.4)	(12.5)	(12.2)	

Table 4.2: Misclassification rates when subsetting the training datasets by selected demographics

Data	Misclassification Rate
All data	0.557
Male	0.647
Female	0.417
EA only	0.618
Family History = 1	0.489
Family History = 2	0.56

Table 4.3: Top 20 Univariate Logistic Regression variables

Retrospective			Prospective		
Variable	Coefficient (SE)	Pvalue	Variable	Coefficient (SE)	Pvalue
rs1992248	0.72(0.23)	0.001	rs903085	1.35(0.45)	0.003
rs2194368	0.67(0.23)	0.003	rs1439630	-1.19(0.43)	0.005
rs11768025	0.58(0.22)	0.01	rs295886	-1.09(0.43)	0.011
rs1629159	0.57(0.22)	0.011	rs6794405	-1.05(0.43)	0.014
History of bipolar	-0.57(0.22)	0.011	rs11160586	1.04(0.44)	0.019
rs1410113	-0.57(0.22)	0.011	rs2648120	-0.93(0.42)	0.027
rs6746908	-0.55(0.22)	0.014	rs10878925	0.91(0.42)	0.031
rs1493554	-0.52(0.22)	0.02	rs10989415	0.91(0.42)	0.031
Sex	-0.51(0.22)	0.021	rs12538005	0.92(0.44)	0.035
rs10134704	0.51(0.22)	0.025	History of bipolar	-0.88(0.42)	0.036
rs10820888	-0.49(0.22)	0.029	rs2241382	-0.87(0.42)	0.04
rs7558514	0.49(0.22)	0.031	rs10075058	0.8(0.42)	0.059
rs10944945	0.47(0.22)	0.037	rs2994991	-0.78(0.42)	0.062
rs4949915	0.46(0.22)	0.038	rs10134704	0.77(0.43)	0.072

4.7 Appendix

Table 4.4: Full results of univariate logistic regression

Retrospective			Prospective		
Variable	Coefficient (SE)	Pvalue	Variable	Coefficient (SE)	Pvalue
rs1992248	0.72(0.23)	0.001	rs903085	1.35(0.45)	0.003
rs2194368	0.67(0.23)	0.003	rs1439630	-1.19(0.43)	0.005

Table 4.4 continued from previous page

rs11768025	0.58(0.22)	0.01	rs295886	-1.09(0.43)	0.011
rs1629159	0.57(0.22)	0.011	rs6794405	-1.05(0.43)	0.014
History of bipolar	-0.57(0.22)	0.011	rs11160586	1.04(0.44)	0.019
rs1410113	-0.57(0.22)	0.011	rs2648120	-0.93(0.42)	0.027
rs6746908	-0.55(0.22)	0.014	rs10878925	0.91(0.42)	0.031
rs1493554	-0.52(0.22)	0.02	rs10989415	0.91(0.42)	0.031
Sex	-0.51(0.22)	0.021	rs12538005	0.92(0.44)	0.035
rs10134704	0.51(0.22)	0.025	History of bipolar	-0.88(0.42)	0.036
rs10820888	-0.49(0.22)	0.029	rs2241382	-0.87(0.42)	0.04
rs7558514	0.49(0.22)	0.031	rs10075058	0.8(0.42)	0.059
rs10944945	0.47(0.22)	0.037	rs2994991	-0.78(0.42)	0.062
rs4949915	0.46(0.22)	0.038	rs10134704	0.77(0.43)	0.072
rs9825823	-0.47(0.22)	0.038	rs13326322	0.77(0.43)	0.074
rs6517050	-0.44(0.22)	0.047	rs3943672	0.77(0.43)	0.074
rs1731240	-0.44(0.22)	0.049	rs13178470	0.73(0.42)	0.08
rs10196335	0.42(0.22)	0.06	rs332182	-0.73(0.42)	0.08
rs1706826	0.42(0.22)	0.064	rs5758527	-0.73(0.42)	0.08
rs2044473	0.41(0.22)	0.066	rs7302134	-0.73(0.42)	0.08
rs1159421	0.41(0.22)	0.067	rs1568293	0.74(0.44)	0.09
rs5758527	0.4(0.22)	0.072	rs6793085	0.71(0.42)	0.091
rs8056559	-0.4(0.22)	0.073	rs9867479	-0.71(0.42)	0.091
rs3087897	0.4(0.22)	0.074	rs10162630	0.69(0.42)	0.099
rs6944714	-0.4(0.22)	0.076	rs2661406	0.69(0.42)	0.099
rs749922	-0.39(0.22)	0.077	rs465387	-0.69(0.42)	0.099
rs4904868	-0.39(0.22)	0.079	rs9446461	-0.69(0.42)	0.099
rs332795	-0.39(0.22)	0.081	rs9825823	0.69(0.42)	0.099

Table 4.4 continued from previous page

rs2766543	0.38(0.22)	0.086	rs12551982	0.66(0.42)	0.114
rs1220046	-0.38(0.23)	0.088	rs2597565	-0.66(0.42)	0.114
rs10125195	-0.37(0.22)	0.099	rs11865138	0.66(0.42)	0.115
rs1810738	0.37(0.22)	0.1	rs1352774	-0.64(0.42)	0.129
rs3784253	-0.36(0.22)	0.109	rs4962528	0.63(0.42)	0.129
rs8040516	-0.36(0.22)	0.109	rs7037204	-0.64(0.42)	0.129
rs7631939	-0.34(0.22)	0.127	rs7920624	-0.63(0.42)	0.129
rs11176433	-0.34(0.22)	0.132	rs8040516	-0.63(0.42)	0.129
rs2241382	-0.33(0.22)	0.139	rs9529497	0.64(0.42)	0.129
rs2916226	0.33(0.22)	0.143	rs6807797	0.63(0.42)	0.135
rs6508273	-0.32(0.22)	0.145	rs12985881	-0.62(0.42)	0.141
rs1431977	-0.32(0.22)	0.147	rs10739677	-0.61(0.42)	0.145
rs2845570	0.32(0.22)	0.153	rs1355834	-0.61(0.42)	0.145
rs2588233	0.32(0.22)	0.155	rs25966	-0.61(0.42)	0.145
rs1580355	0.32(0.22)	0.157	rs8056559	0.64(0.44)	0.149
rs9381469	-0.31(0.22)	0.166	rs11633606	0.59(0.42)	0.158
rs3850143	0.3(0.22)	0.179	rs4904868	0.59(0.42)	0.158
rs10783287	0.29(0.22)	0.188	rs5024581	0.59(0.42)	0.158
rs9867479	-0.29(0.22)	0.197	rs593479	0.59(0.42)	0.158
rs1477242	-0.28(0.22)	0.207	rs1823803	-0.58(0.42)	0.162
rs6793085	-0.28(0.22)	0.208	rs2938786	0.59(0.43)	0.169
rs593479	0.28(0.22)	0.214	rs17405754	0.59(0.44)	0.173
rs11249523	-0.28(0.23)	0.214	rs10775439	0.56(0.42)	0.178
rs354153	-0.27(0.22)	0.217	rs1220046	0.56(0.42)	0.178
rs2194573	0.27(0.22)	0.22	rs6517050	-0.56(0.42)	0.178
rs2203859	0.27(0.22)	0.227	rs181586	-0.56(0.42)	0.179
rs10981316	0.27(0.22)	0.232	rs2287630	-0.56(0.42)	0.179
rs716513	0.27(0.22)	0.232	rs1671471	0.55(0.42)	0.191

Table 4.4 continued from previous page

rs7419684	-0.26(0.22)	0.24	rs652105	0.55(0.42)	0.191
rs2436104	0.26(0.22)	0.241	rs8006194	0.55(0.42)	0.191
rs4910146	0.26(0.22)	0.241	rs7444222	-0.54(0.42)	0.198
rs2033859	-0.26(0.22)	0.248	rs4678492	-0.53(0.41)	0.199
rs12453291	-0.26(0.22)	0.251	rs7573382	-0.52(0.42)	0.215
rs4236814	-0.25(0.22)	0.256	rs3850143	-0.51(0.41)	0.221
rs4901203	-0.25(0.22)	0.256	rs588789	0.51(0.41)	0.221
rs10868796	-0.25(0.22)	0.259	rs6465658	0.51(0.41)	0.221
rs3125289	0.25(0.22)	0.265	rs2242259	0.51(0.42)	0.222
rs8006194	-0.25(0.22)	0.269	AA	0.8(0.68)	0.24
rs10062387	-0.24(0.22)	0.272	rs7006331	0.49(0.42)	0.24
rs447107	0.24(0.22)	0.273	rs2194368	-0.48(0.42)	0.249
rs11160586	0.24(0.22)	0.278	rs4700135	-0.48(0.42)	0.255
rs6018076	-0.24(0.22)	0.278	rs332795	0.48(0.42)	0.258
rs7114464	0.24(0.22)	0.279	rs7992637	0.49(0.43)	0.259
rs9470973	0.24(0.22)	0.279	rs1940658	-0.46(0.41)	0.266
rs2188584	-0.24(0.22)	0.279	rs1580355	-0.46(0.42)	0.268
rs17315876	0.24(0.22)	0.292	rs1262940	0.45(0.42)	0.285
rs181586	0.23(0.22)	0.295	rs1779549	-0.45(0.42)	0.285
rs6807797	0.23(0.22)	0.3	rs6826171	0.45(0.42)	0.285
rs1671471	0.23(0.22)	0.31	rs10944945	-0.43(0.41)	0.294
rs10858494	-0.23(0.22)	0.31	rs6508273	0.43(0.41)	0.294
rs857633	-0.23(0.22)	0.31	rs9949868	-0.43(0.41)	0.294
rs9545650	0.22(0.22)	0.313	rs9998008	-0.43(0.41)	0.294
rs1355834	-0.22(0.22)	0.314	rs10981316	0.44(0.43)	0.299
rs332182	-0.22(0.22)	0.324	rs193689	0.44(0.43)	0.299
rs7604877	0.22(0.22)	0.333	rs12033696	0.42(0.41)	0.315
rs6498053	-0.22(0.22)	0.333	rs6972063	0.42(0.41)	0.315

Table 4.4 continued from previous page

rs13006508	-0.21(0.22)	0.338	rs17692528	-0.41(0.42)	0.32
rs10782506	0.21(0.22)	0.344	rs1241927	-0.41(0.41)	0.322
rs9288053	0.21(0.22)	0.348	rs12507758	0.41(0.41)	0.322
rs3736919	-0.21(0.22)	0.353	rs1753430	-0.41(0.41)	0.322
rs1405879	0.21(0.22)	0.357	rs677476	0.4(0.42)	0.33
rs100224	-0.2(0.22)	0.36	rs2542242	0.41(0.42)	0.333
rs1823803	0.2(0.22)	0.363	rs10125195	-0.39(0.41)	0.348
rs9487881	0.2(0.22)	0.38	rs749922	-0.39(0.41)	0.348
rs1262940	-0.19(0.22)	0.382	rs7558514	-0.39(0.41)	0.348
rs10277195	0.19(0.22)	0.383	rs977869	0.39(0.41)	0.348
rs3857152	-0.19(0.22)	0.383	rs6599645	0.39(0.42)	0.352
rs10794720	0.27(0.32)	0.39	rs7842027	-0.39(0.42)	0.352
rs2597565	0.19(0.22)	0.396	rs720959	0.37(0.42)	0.369
rs25966	-0.19(0.22)	0.402	rs1836868	-0.37(0.42)	0.371
rs4839259	-0.19(0.22)	0.405	rs2893735	0.37(0.42)	0.381
rs1940658	-0.18(0.22)	0.41	rs2845570	0.34(0.41)	0.406
rs7573382	-0.18(0.22)	0.415	rs11782265	0.34(0.41)	0.415
rs2466384	-0.18(0.22)	0.42	rs13336069	-0.34(0.41)	0.415
rs465387	-0.18(0.22)	0.421	rs4549499	0.33(0.41)	0.424
rs11143486	-0.18(0.22)	0.424	rs1455501	0.33(0.42)	0.428
rs10775504	0.18(0.22)	0.43	rs7114097	-0.33(0.41)	0.429
rs2445587	0.17(0.22)	0.432	rs2466384	0.32(0.41)	0.443
rs2154119	0.17(0.22)	0.44	rs306207	0.32(0.41)	0.443
rs11865138	-0.17(0.22)	0.44	rs9353722	0.32(0.41)	0.443
rs7006331	0.17(0.22)	0.444	rs9288053	-0.31(0.41)	0.449
rs2893735	-0.17(0.22)	0.456	rs4239162	-0.31(0.41)	0.451
rs3910188	0.16(0.22)	0.459	rs4241340	0.31(0.41)	0.451
rs7992637	0.16(0.22)	0.467	rs6774729	0.31(0.41)	0.458

Table 4.4 continued from previous page

rs1921032	0.16(0.22)	0.47	rs11744698	0.3(0.42)	0.468
rs13060099	-0.16(0.22)	0.477	rs7186479	0.3(0.42)	0.468
rs4700135	-0.15(0.22)	0.493	rs4945362	-0.3(0.42)	0.478
rs4678492	0.15(0.22)	0.502	rs11523187	0.29(0.41)	0.482
rs2242259	0.15(0.22)	0.506	rs1340389	0.29(0.41)	0.482
rs4238558	0.15(0.22)	0.506	rs4975784	0.29(0.41)	0.482
rs295886	-0.15(0.22)	0.506	rs7195440	-0.29(0.41)	0.482
rs4975784	-0.15(0.22)	0.508	rs11768025	0.3(0.43)	0.488
rs10162630	-0.15(0.22)	0.511	rs1431977	-0.28(0.42)	0.494
rs1439630	-0.15(0.22)	0.511	rs9545650	0.28(0.42)	0.497
rs1340389	-0.14(0.22)	0.518	rs9487881	0.28(0.42)	0.501
rs6826171	0.14(0.22)	0.519	rs10263303	-0.26(0.41)	0.522
Other	-0.33(0.52)	0.525	rs1405879	0.26(0.41)	0.522
rs7516478	0.14(0.22)	0.527	rs2722276	-0.26(0.41)	0.522
rs10778029	0.14(0.22)	0.53	rs3910188	-0.26(0.41)	0.522
rs12033696	-0.14(0.22)	0.543	rs9381469	0.26(0.41)	0.522
rs2648120	0.13(0.22)	0.548	rs9470973	-0.26(0.42)	0.534
rs1693571	0.13(0.22)	0.557	rs13060099	0.26(0.41)	0.535
rs6993722	-0.13(0.22)	0.558	rs11683590	0.25(0.41)	0.539
rs1926502	0.13(0.22)	0.563	rs10775504	0.25(0.41)	0.553
rs9446461	0.13(0.22)	0.564	rs1410113	0.25(0.41)	0.553
rs6903615	0.13(0.22)	0.568	rs6095134	-0.25(0.41)	0.553
rs4239162	-0.13(0.22)	0.568	rs6746908	0.25(0.41)	0.553
rs9949868	-0.13(0.22)	0.568	rs7678054	0.25(0.41)	0.553
rs978599	-0.13(0.22)	0.574	rs1706826	-0.24(0.41)	0.563
rs7444222	0.12(0.22)	0.579	rs928768	0.24(0.41)	0.563
rs10456458	-0.12(0.22)	0.581	rs7142620	-0.23(0.42)	0.574
rs1885396	-0.12(0.22)	0.584	rs12196685	-0.23(0.42)	0.58

Table 4.4 continued from previous page

rs12929924	-0.12(0.22)	0.585	rs1926502	0.23(0.42)	0.58
rs10752019	0.12(0.22)	0.586	rs6993722	0.23(0.42)	0.58
rs11782265	0.12(0.22)	0.586	rs9898391	-0.23(0.42)	0.58
rs5019664	0.12(0.22)	0.595	rs2154119	-0.23(0.42)	0.59
rs13178470	-0.12(0.22)	0.596	rs1159421	-0.23(0.42)	0.593
rs4945362	-0.12(0.22)	0.6	rs100224	0.22(0.41)	0.596
rs1568293	-0.12(0.22)	0.605	rs11263943	0.22(0.41)	0.596
rs1352774	0.11(0.22)	0.606	rs13006508	-0.22(0.41)	0.596
rs652105	0.11(0.22)	0.606	rs1731240	0.22(0.41)	0.596
rs13326322	-0.11(0.22)	0.61	rs2436104	-0.22(0.41)	0.596
rs6557416	-0.11(0.22)	0.612	rs857633	0.22(0.41)	0.596
rs4241340	0.11(0.22)	0.619	rs4901203	0.22(0.42)	0.6
rs720959	-0.1(0.22)	0.638	rs1877818	-0.21(0.41)	0.605
rs193689	0.1(0.22)	0.653	rs6903615	-0.22(0.43)	0.605
rs4679742	-0.1(0.22)	0.653	rs4839259	-0.21(0.41)	0.609
rs6774729	0.1(0.22)	0.664	rs2100076	0.21(0.41)	0.61
rs7037204	-0.1(0.22)	0.667	rs12929924	0.21(0.42)	0.62
AA	-0.29(0.68)	0.667	rs185122	-0.2(0.42)	0.627
rs7842027	-0.09(0.22)	0.672	rs2386123	0.2(0.42)	0.627
rs7142620	-0.09(0.22)	0.677	rs3125289	-0.2(0.42)	0.627
rs6950504	-0.09(0.22)	0.678	rs4679742	-0.2(0.42)	0.627
rs12196685	-0.09(0.22)	0.679	rs10147990	0.2(0.43)	0.639
rs2287630	0.09(0.22)	0.682	rs10820888	-0.19(0.41)	0.641
rs717908	0.09(0.22)	0.682	rs2588233	0.19(0.41)	0.641
rs1528779	0.09(0.22)	0.687	rs17315876	-0.19(0.42)	0.648
rs2722276	0.09(0.22)	0.689	rs933746	0.19(0.42)	0.65
rs7302134	-0.09(0.22)	0.69	rs1286648	0.18(0.41)	0.657
HNA	-0.22(0.57)	0.693	rs2916226	0.17(0.41)	0.674

Table 4.4 continued from previous page

rs13409819	0.09(0.22)	0.696	rs7114464	-0.18(0.43)	0.674
rs2386123	0.09(0.22)	0.697	rs761130	0.17(0.41)	0.674
rs2049306	0.09(0.22)	0.701	rs11176433	-0.17(0.41)	0.687
rs6859309	-0.09(0.22)	0.701	rs12453291	0.17(0.41)	0.687
rs17692528	0.08(0.22)	0.709	rs2050083	-0.17(0.41)	0.687
rs1455501	0.08(0.22)	0.71	rs2198010	-0.17(0.42)	0.693
rs7920624	0.08(0.22)	0.712	rs1958497	0.16(0.42)	0.701
rs2642990	-0.08(0.22)	0.715	rs958249	0.16(0.42)	0.705
rs2542242	-0.08(0.22)	0.727	rs1921032	-0.15(0.41)	0.722
rs1836868	-0.08(0.22)	0.728	rs7419684	-0.15(0.41)	0.722
rs10739677	0.08(0.22)	0.73	rs4899563	-0.14(0.41)	0.733
rs2661406	-0.08(0.22)	0.732	rs10778029	0.15(0.43)	0.734
rs9998008	-0.08(0.22)	0.732	rs4236814	0.13(0.42)	0.755
rs10989415	-0.07(0.22)	0.737	rs10794720	-0.18(0.6)	0.761
rs2274064	-0.07(0.22)	0.742	rs10277195	-0.12(0.41)	0.77
rs4962528	-0.07(0.22)	0.742	rs1574249	-0.12(0.41)	0.77
rs6095134	-0.07(0.22)	0.744	rs2642990	0.12(0.41)	0.77
rs11633606	0.07(0.22)	0.746	rs2766543	0.12(0.41)	0.77
rs10878925	0.07(0.22)	0.752	rs6498053	-0.12(0.41)	0.77
rs11071215	-0.07(0.22)	0.757	rs11249523	-0.12(0.42)	0.78
rs6794405	0.07(0.22)	0.767	rs1992248	-0.12(0.42)	0.78
rs185122	-0.06(0.22)	0.793	rs10456458	-0.11(0.41)	0.787
rs9359520	-0.06(0.22)	0.796	rs4949915	-0.12(0.43)	0.787
rs6680193	-0.06(0.22)	0.798	rs1885396	-0.11(0.42)	0.795
rs6972063	0.05(0.22)	0.814	rs6018076	0.11(0.43)	0.803
rs1877818	-0.05(0.22)	0.814	rs13028997	0.1(0.41)	0.805
rs11744698	0.05(0.22)	0.819	rs4975957	0.1(0.41)	0.805
rs12507758	-0.05(0.22)	0.819	rs717908	-0.1(0.41)	0.805

Table 4.4 continued from previous page

rs4855980	-0.05(0.22)	0.819	rs2203859	-0.09(0.41)	0.819
rs10075058	-0.05(0.22)	0.823	rs4324759	-0.09(0.41)	0.819
rs4899563	0.05(0.22)	0.825	rs6859309	-0.09(0.42)	0.832
rs10774707	-0.05(0.22)	0.825	rs6944714	-0.09(0.42)	0.832
rs1779549	-0.05(0.22)	0.83	rs1528779	0.09(0.42)	0.839
rs10147990	0.05(0.22)	0.832	rs2044473	0.09(0.42)	0.839
rs928768	0.05(0.22)	0.835	rs4910146	0.09(0.42)	0.839
rs12551982	-0.05(0.22)	0.835	rs6680193	0.09(0.42)	0.839
rs4774583	0.05(0.23)	0.835	rs7604877	-0.08(0.41)	0.855
rs9353722	0.04(0.22)	0.84	rs11143486	-0.07(0.41)	0.868
rs5024581	-0.04(0.22)	0.846	rs1693571	-0.07(0.41)	0.868
rs9898391	-0.04(0.22)	0.849	rs4238558	-0.07(0.41)	0.868
rs11263943	-0.04(0.22)	0.853	rs805007	0.07(0.41)	0.868
rs13028997	0.04(0.22)	0.854	rs2194573	-0.07(0.41)	0.871
rs3943672	-0.04(0.22)	0.856	HNA	-0.1(0.68)	0.878
rs9529497	0.04(0.22)	0.861	rs6950504	0.06(0.41)	0.878
rs265518	0.04(0.22)	0.867	rs1629159	-0.06(0.42)	0.879
rs2810876	-0.04(0.22)	0.87	rs3784253	0.06(0.42)	0.884
rs13336069	-0.04(0.22)	0.872	rs9359520	0.06(0.41)	0.884
rs2198010	-0.03(0.22)	0.875	rs10783287	-0.06(0.42)	0.891
rs1958497	0.03(0.22)	0.876	rs10868796	-0.06(0.42)	0.891
rs903085	0.03(0.22)	0.878	rs2810876	0.06(0.42)	0.891
rs3829382	-0.03(0.22)	0.882	rs4774583	0.06(0.42)	0.891
rs4549499	0.03(0.22)	0.883	rs10062387	-0.05(0.41)	0.906
rs11523187	0.03(0.22)	0.887	rs10752019	-0.05(0.41)	0.906
rs10263303	-0.03(0.22)	0.887	rs2188584	0.05(0.41)	0.906
rs2938786	0.03(0.22)	0.89	rs2791603	0.05(0.41)	0.906
rs4324759	0.03(0.22)	0.893	rs10196335	0.04(0.42)	0.917

Table 4.4 continued from previous page

rs1241927	0.02(0.22)	0.911	rs6557416	-0.04(0.42)	0.922
rs4975957	-0.02(0.22)	0.914	rs354153	-0.03(0.41)	0.938
rs6465658	-0.02(0.22)	0.914	rs2049306	-0.03(0.41)	0.942
rs933746	-0.02(0.22)	0.914	rs265518	0.03(0.41)	0.942
rs977869	0.02(0.22)	0.916	rs5019664	0.03(0.41)	0.942
rs2050083	0.02(0.22)	0.925	rs7516478	-0.03(0.41)	0.942
rs10775439	-0.02(0.22)	0.925	rs10782506	0.02(0.41)	0.956
rs1574249	-0.02(0.22)	0.931	rs10858494	0.02(0.41)	0.956
rs1775715	0.02(0.22)	0.936	rs11071215	-0.02(0.41)	0.956
rs1286648	-0.02(0.22)	0.938	rs1810738	0.02(0.41)	0.956
rs11683590	0.02(0.22)	0.939	rs2274064	0.02(0.41)	0.956
rs12985881	-0.02(0.22)	0.941	rs2445587	0.02(0.41)	0.956
rs2100076	0.02(0.22)	0.946	rs3087897	0.02(0.41)	0.956
rs588789	0.01(0.22)	0.947	rs447107	0.02(0.41)	0.956
rs805007	-0.01(0.22)	0.948	rs4855980	-0.02(0.41)	0.956
rs7186479	-0.01(0.22)	0.956	rs716513	-0.02(0.41)	0.956
rs2791603	-0.01(0.22)	0.963	rs978599	0.02(0.41)	0.956
rs12538005	0.01(0.22)	0.967	Sex	-0.03(0.63)	0.964
rs7678054	0.01(0.22)	0.967	rs3857152	0.02(0.42)	0.969
rs6599645	-0.01(0.22)	0.967	rs10774707	0.01(0.41)	0.972
rs761130	-0.01(0.22)	0.973	Other	0.03(0.79)	0.974
rs958249	0.01(0.22)	0.978	rs1493554	0.01(0.41)	0.974
rs677476	-0.01(0.22)	0.978	rs1477242	-0.01(0.41)	0.983
rs7114097	-0.01(0.22)	0.978	Easia	-15.29(1029.12)	0.988
rs1753430	0(0.22)	0.984	rs13409819	0(0.41)	0.993
Easia	-15.65(840.27)	0.985	rs1775715	0(0.41)	0.993
rs306207	0(0.22)	0.986	rs2033859	0(0.41)	0.993
rs17405754	0(0.22)	0.995	rs3736919	0(0.41)	0.993

Table 4.4 continued from previous page

rs2994991	0(0.22)	0.995	rs3829382	0(0.41)	0.993
rs7195440	0(0.22)	0.995	rs7631939	0(0.41)	0.993

CHAPTER 5

Comparison of Factors Associated with Quitting Smoking in Health Care Workers Using Fuzzy Forests While Adjusting for Self-response Weights

5.1 Abstract

5.1.1 Background:

Smoking is the leading preventable cause of death in the US. It is important to understand current everyday smokers to help guide their attempts to quit smoking. Health care providers are a critical part of educating smokers to quit smoking, but they, despite their first-hand knowledge of the harms of smoking, are not immune to become smokers. Healthcare providers who smoke may be a barrier to smoking cessation in their patients, thus it is important to understand the determinants of current everyday smokers in health care professionals.

5.1.2 Methods:

We utilize the Tobacco Use Supplement of the Current Population survey subset to health care professionals who are current everyday smokers. We further subdivide this subset

into Diehards - those who do not wish to quit smoking and Tryhards - those that wish to quit smoking. We use both machine learning and traditional methods to find the predictors of being a Diehard or a Tryhard.

5.1.3 Results:

We find that it is necessary to use the survey design weights in machine learning, even for variable selection. Predictors of being a Diehard are permissive smoking rules such as it being allowed anywhere in the household and long-term everyday smoking. For Tryhards we find that smoking being allowed in bars and clubs as well as smoking being allowed in the home as having lower odds of being a Tryhard versus a ban on smoking in these areas. This suggests that policy changes, such as smoking bans, could be helpful in current everyday smokers successfully quitting.

5.1.4 Conclusion:

Machine learning is not a substitute for careful analysis whereby the survey design is taken into account. For health care providers, for whom knowledge campaigns might not be effective to reduce smoking rates as they already have seen firsthand the detrimental effects of smoking, different legislation may be in order. Based on the results of the survey, full smoking bans that could reduce the social desirability of smoking may help these people successfully quit smoking.

5.2 Introduction

According to the Centers for Disease Control and Prevention, cigarette smoking is responsible for at least 480,000 deaths every year and more than 16 million Americans live with a chronic smoking-related disease ([57]). Smoking is the number one cause of preventable death in the United States. Despite numerous public health campaigns and abundant evidence linking smoking to cancers and other serious health problems, in 2015, there still remain 15.1% of the adults in the US who are smokers, with 11.4% who smoke daily [45].

Because of the public health concerns related to smoking, along with implementing interventions on the person-level, instituting policy changes on a state-level can also be an effective means to reduce the prevalence of smoking. The U.S. Centers for Disease Control and Prevention (CDC) reports that in 2015, the majority of smokers wish to quit, with 68.0% of adult smokers reported as wanting to stop smoking, 55.4% having made a quit attempt in the last year, and 7.4% reported recently having quit smoking [16]. In 2017, current smoking declined from 20.9% in 2005 to 14% in 2017 [82]. Statewide policy changes have been implemented across the US to help curb smoking and prevent second hand smoke. While at the discretion of each state, as of March 2019, 28 states have instituted laws that prohibit smoking in enclosed workplaces including bars and restaurant restaurants [3]. In addition to public areas across the US, restrictions are also placed on public housing. As of July 31, 2018, the Department of Housing and Urban Development has required that all public housing agencies and multifamily federally assisted properties have a smoke-free policy for these properties [1].

National surveys, such as the Tobacco Use Supplement, which is incorporated in the Current Population Survey (TUS-CPS) collect large quantities of information from individuals across the entire US [4]. The TUS-CPS collects information on current cigarette smoking status, smoking history, amount spent on cigarettes, and attitudes toward smoking policies

along with other smoking-related topics [5].

In general, smokers are more likely to be male, younger, have lower educational attainment, lower annual household income, more likely to live in the Midwest versus the west, higher among divorced/separated or widowed versus married or living with a partner, more likely to be uninsured or on Medicaid [82].

It is interesting to note, that even though individuals that work in health care roles and can be expected to be knowledgeable about the devastating health consequences of smoking, they are not immune from the draws of smoking, as the prevalence of current smokers among health care and social assistance workers is around 16% [76]. In particular, it was found that licensed practical nurses (LPNs) have the highest rate of smoking (20.55%) and were less likely to quit smoking during that same period [65]. Medical care professionals who smoke can be seen as a barrier to quit smoking among their patients [65]. As such, it is important to understand the reasons for smoking, barrier to quitting, and possible interventions for this sub-population of smokers. Among the subset of health care professionals, we focused in particular on those who are current everyday smokers. Of the current everyday smokers, We assessed subjects who are interested in quitting smoking (Tryhards) and those that have expressly stated that they are not interested in quitting smoking (Diehards). The resulting analysis, will allow for a wider set of factors upon which to explore in future interventions and aid in cessation attempts.

5.3 Data

We obtained publicly available data from the Tobacco Use Supplement of the Current Population Survey (TUS-CPS). The TUS-CPS uses stratified probability sampling to provide representative estimates of the population by occupation and has been administered since 1992 with data being collected every 3-4 years [4]. For this analysis, we use the 2010-

2011 data administered to those 18 years or older. The data includes 160 replicate weights and 721 original variables and additional variables, which included some composite variables of related items asked in the questionnaire. Some of these 721 variables included in the questionnaire inquired about current smoking status, amount smoked, use of menthol cigarettes, smoking history, level of nicotine dependence, and cost of cigarettes. The data was subset to include only those in health care related occupations, such as dentist, pharmacists, nurses, therapists that are current every day smokers (answered do you now smoke cigarettes everyday, some days, or not at all and responded everyday or some days) thus retaining only 876 individuals out of the original 227,722 to be used in subsequent analyses.

Some items and individuals were removed prior to analysis. Variables removed included those that were useful for the study but not necessarily interpretable, such as allocation weight for the variables or line numbers of a question, questions not asked of those that were current everyday smokers, variables used in the formation of composite variables, or were a sampling weight and were not used in the subsequent weighting of the data, along with those missing at least 23% of the observations. Thus, after removal of these variables, there are only 99 potential covariates to be analyzed on the 876 remaining individuals. The relationships among these 99 variables are explored to determine their relationship to their willingness to quit smoking.

A persons willingness to quit smoking allows them to be categorized as a Diehard or a Tryhard. A Diehard is anyone who indicated that they had not stopped smoking for one day or longer in the past 12 months because they were trying to quit, and in fact had never made a serious attempt to stop smoking even for a day, and also that they did not indicate seriously considering quitting smoking within the next 6 months, and also that they were not interested in quitting smoking determined by a score of 7 or less out of 10 for their interest in quitting. A Tryhard is a non-Diehard individual that is very interested in quitting smoking, indicating an 8 or higher out of 10 in their interest. Of the 876 current everyday smokers, 223 are Diehards and 272 are Tryhards.

5.4 Methods

The TUS-CPS, collects numerous pieces of information that might involve a nested data structure where subcategories of the levels of a particular variable might differ across the different levels of that variable. For example: a general question such as work status might have response options distinguishing full-time, at least 50% part-time, less than 50% part-time, or no work while variables might assess when applicable the reason for not working full-time, which would imply structural missingness for those who do work full-time. Another question might gauge how many hours a week a person works, and yet another might gauge whether the individual has more than one occupation. With nested questions, strong associations between measured variables can emerge from data collected using nested questions. Further, many surveys ask similar questions but in different ways to ascertain respondent consistency, inducing correlation as well. In exploring the relationship between being a Diehard smoker (or a Tryhard) and other variables collected in the TUS-CPS survey, logistic regression would commonly be utilized. However due to the sheer quantity of information collected from each person in the survey, along with the correlation among the variables, using logistic regression can be challenging. When dealing with high-dimensional data, especially in the case where the number of parameters (variables) exceeds the number of observations (respondents) one can consider machine learning techniques to identify important parameters and aid with variable selection. Machine learning techniques, such as Random Forests, can quickly identify important variables, including those that may go unnoticed among the abundance of items collected.

Random Forests is a popular machine learning technique that is fast, computationally efficient in the context of the small-n / large-p problem for variable selection, and has been shown to have good predictive ability. However, in the presence of correlated variables, multiple investigations have shown that the resulting variable importance rankings of the variables are biased ([74],[54], [52], [15], [38]). Fuzzy Forests is an extension of Random Forests that has all of the beneficial characteristics that Random Forests have, notably being

non-parametric and being able to handle large numbers of variables including various data types, but Fuzzy Forests can also account for groups of correlated variables. Fuzzy forests and Random Forests do not produce estimates or p-values and do not usually incorporate sampling weights into the algorithm; rather the methods yield fitted models that satisfy an estimation criterion such as optimizing a measure of predictive accuracy. We will run Fuzzy Forests two ways, one without the survey sample weights and one with the sample weights to assess the role of weighting in variable importance selection. In this project, we use the weighted Fuzzy Forests to assess the important variables associated being a Diehard and separately for Tryhards. The results are then used to perform the more traditional survey-weighted logistic regression to obtain parameter estimates.

Fuzzy Forests and Random Forests are both ensemble classifiers that combine the results of many binary decision trees. They both create multiple trees based off of a single dataset by taking multiple bootstrap samples with replacement from the data. Each tree is built off of one of these bootstrap samples. The number of bootstrap samples and hence the number of trees in the forest are controlled by the tuning parameter ntree. The bootstrap samples are taken with replacement, thus each tree is actually built on approximately 66% of the entire sample. Since each tree does not utilize all the data, an out-of-bag (OOB) error rate based on the unused data can be obtained for each tree. These OOB samples can be used as a test set to obtain predictions from each tree that was not grown using that data. The resulting prediction for each observation averages the prediction across all the trees. Similarly, the error rate obtained from the OOB samples are averaged to obtain the OOB error rate. Along with taking samples from the data upon which to grow each tree, within each tree, at each node, the best variable upon which to split the data is chosen from a random subset of the variables. The results from the trees are combined to provide a ranking of all the variables in terms of how important they are in predicting the outcome.

However, unlike Random Forests, Fuzzy Forests is able to model correlational aspects in the data. This is done via a two-step procedure. Fuzzy Forests first separates the data into

sets of similarly correlated variables. This can be achieved via weighted correlation network analysis or by user specification. Then recursive feature elimination Random Forests are performed on each of the correlated sets or modules. In the case of the TUS-CPS data, it is simpler to specify candidate sets of variables to be the focus of the analysis. Among the variables, four groups or modules of variables were chosen; specifically job/finance questions, smoking questions, demographics characteristics, or location/region indicators. Each set of trees was grown using only variables found in each module. From the resulting variable importance list, the bottom set of variables was removed until the OOB error rate attained its lowest observed value. This process was repeated for each group of variables. Once the top list of variables is obtained from each group, the process is again repeated using all the variables that were retained, thus providing a top set of variables related to the outcome.

The present analysis uses weighted bootstrap samples in the Fuzzy Forests algorithm to adjust for the sampling design used in the collection of the TUS-CPS data. The weight used in the weighted bootstrap is the self-response weight from the survey, which included a non-interview adjustment and a self-response adjustment. For the Fuzzy Forest algorithm, the average of five weighted bootstrap samples was used to illustrate the effect of incorporating a survey weight in the analysis. The average ranking of the top 20 variables from each weighted bootstrap sample was used to give a measure of variable importance along with the average OOB error rate.

The weighted and unweighted Fuzzy Forests results were then used in a weighted logistic regression. Logistic regression is a familiar technique for most researchers. While machine learning and Fuzzy Forests in particular, are arguably better suited for this type of exploratory analysis, many researchers who are unfamiliar with these methods might opt to use logistic regression. However, convergence issues due to the quantity of variables and the correlation among the predictors can prevent a logistic regression from using all the variables at once. On the other hand, utilizing a weighted logistic regression building on the results of a machine learning algorithm allows for all the variables to be explored

while producing results that will be familiar to researchers accustomed to using parametric methods to facilitate both convergence of the estimation procedure and interpretation of the results. For the weighted logistic regression, a balanced repeated replication design is used with a Fays correction of 0.5 along with the self-response weights and the replicate weights to determine if any of the variables are significantly related to whether a person is a Diehard or a Tryhard.

At the stage of performing logistic-regression analysis, we collapsed a number of variables that had multiple levels. Selected variables that were collapsed include professions, which we created an indicator of nursing versus other; for marital status, married versus not married. All statistical analyses were performed using SAS (SAS Institute Cary, NC), version 9.3, and R Studio 1.1.463 using the fuzzyforest library.

As in any data-analysis setting, the definitions of variables have implications for the extent to which covariates in a prediction model are correlated with one another, the approach taken here anticipating that some users will enter covariates into an analysis without preprocessing and without much attention to the multicollinearity, is to enter variables in the analysis close to the way they were collected with only limited preprocessing. In doing so, we are not recommending that preprocessing is to be avoided; rather we are interested in how various machine-learning algorithms perform in such contexts.

5.5 Results

The demographic characteristics of the 2010-2011 survey data from 876 health care workers who are current everyday smokers are presented in Table 1. On average, the health-care workers are around 40 years old, with a plurality being married (36%) and female (87%). Note that this differs from the general population where smokers are disproportionately male and either separated/divorced/widowed. The professional group that is most

represented in the sample is nursing, psychiatric or home health care aides (37%). This is consistent with previous studies that have found that licensed practical nurses had the highest rates of smoking prevalence among the health care industry and that physicians had the lowest prevalence rate [64]. Among this sample of individuals in the health care industry who are current everyday smokers, 26% are Diehard smokers who have no interest in quitting smoking as compared to the Tryhards that make up approximately 30%; the remaining individuals belong to neither of these two groups. Table 2 lists these same demographic features separated by whether the subject identifies as being a Diehard or a Tryhard. The Diehards have a larger percentage of nurses, which include registered nurses, nurse anesthetists and nursing/psychiatric/home health aids as compared to Tryhards (58% vs. 45%).

Naively, an unweighted Fuzzy Forests analysis was performed. It is common that this type of analysis is performed in a machine-learning context where the survey structure and the survey weights are not utilized. We thought there might be different variables that influence someone who does not want to quit smoking versus someone who does; consequently, a separate forest was constructed to predict who was likely to be a Diehard and who is likely to be a Tryhard. The goal of the Fuzzy Forests analysis is to list the most important variables predicting the outcome. The initial analyses of Diehards using the unweighted data results in an average OOB error rate of 31.51%, and the average OOB error rate for Tryhards is 34.82%. Of the variables included in the predictor space, the highest ranking fifteen variables are presented in Table 3. We list the variables that each analysis had in common as well as those that were different in each analysis. Common variables included how many hours the individuals worked, the number of years they smoked every day, age, occupation, full-time or part-time job status, number of cigarettes smoked each day, and how many hours they worked at their main job. Diehards had unique predictors such as a medical doctor telling the person to quit smoking, years of every-day smoking, educational attainment and location. For Tryhards policy questions about smoking in bars was important, as well as average number of daily cigarettes, hours worked, menthol cigarettes and if the person bough cigarettes by the carton or the pack.

Comparatively, when the self-response weights are included, the top 20 variable rankings also included many common variables between the Diehards and Tryhards. Table 4 lists the top variables in terms of importance for the implementation of Fuzzy Forests that includes the survey weights. Some such common variables are the time until first cigarette, age, type of job, and presence of children. However, its the unique variables that may better predict if one will be a Diehard or a Tryhard. Interestingly, among the Tryhards, it is allowance of smoking within bars, clubs and cocktail lounges that is especially important in determining whether a person will be a Tryhard or not. The weighted Fuzzy Forests implementation also identified other variables that unweighted Fuzzy Forests did not identify, such as smoking rules inside the home and having children. The weighted Fuzzy Forests implementation had better performance in terms of AUC and predictive error than unweighted Fuzzy Forests.

Fuzzy Forests can also be used for feature selection to include in further analysis. In this case, survey-weighted logistic model was performed as a subsequent analysis strategy. The results of the model are presented in Table 5. Figure 1 shows the variable importance plots for Diehards. Weighted logistic regression, after controlling for age of first use of cigarettes, gender and race, indicated that important predictors of whether someone will be a Diehard smoker or not, include the number of cigarettes used daily, how long the individual has smoked every day, and whether smoking is allowed in the home. The odds of being a Diehard are significantly higher among those who allow smoking anywhere in the home compared to those who do not allow smoking anywhere in their home ($p= 0.01$). The odds of being a Diehard is lower for those who do not smoke every day as compared to those that do smoke every day ($p = 0.0005$). Also, after controlling for the other variables in the model, the odds of being a Diehard are lower as the number of cigarettes smoked on average increases ($p = 0.002$).

Among the Tryhards, the survey-weighted logistic regression also yielded similar results. The results are presented in Table 6, with Figure 2 presenting the variable importance

plot for Tryhards. Similarly, after controlling for age of first use of cigarettes, race and gender, not allowing smoking in bars and clubs as well as not being allowed to smoke anywhere in the home are associated with being a Tryhard. If smoking is allowed anywhere in the home, the odds of being a Tryhard are lower than if it is not allowed anywhere ($p = 0.02$). Similarly, if smoking is allowed in some ($p=0.04$) or all ($p=0.05$) places in bars, cocktail lounges or clubs, then the odds of being a Tryhard are lower than in places where smoking is not allowed.

5.6 Discussion

Fuzzy Forests with survey weights was observed to have better predictive performance compared to ignoring the survey design, with an average AUC of around 95%. This suggests the importance of using survey weights even when doing variable screening and selection. Fuzzy Forests found that profession was an important predictor of smoking behavior, operationalized here as being either Diehard or Tryhard. This finding is consistent with other surveys which found that Licensed Practical Nurses are more likely to be smokers than physicians.

As noted earlier, we characterized smoking behavior of every day current smokers into Diehards (those who do not wish to quit smoking) and Tryhards (those who have a stated high desire to quit smoking). We found in weighted survey logistic regression analysis that current every day smokers who smoke every day for nearly all the years that they have smoked are more likely to be a Diehard. Diehards are also more likely to be allowed to smoke anywhere in their homes, with people who can smoke anywhere in their home, being almost twice as likely to be a Diehard as when smoking is not allowed in the home. Diehards also report lower numbers of maximum daily cigarettes than non-Diehards. Perhaps, akin to addictions where individuals do not perceive their own substance use as excessive, there might be a perception that smoking is not a problem. It is important to understand this

group, as Diehard smokers present a challenge to reach to try to change their behavior.

The examination of Tryhards resulted in some interesting policy implications. Smoking being allowed in bars and clubs was associated with lower odds of being a Tryhard versus it being completely banned. Currently, 28 states have comprehensive smoking bans that include smoking in bars and lounges. Studies have shown that anti-smoking legislation can spur smoking cessation attempts [24].

Cessation efforts may be due, at least in part, to the way that smoking bans reduce the social desirability of smoking and can lead to less socially cued smoking. This may be especially true in bars and lounges, as many people who report being a social smoker report that they only smoke in such settings. Thus banning smoking in bars and lounges may lead these people to quit. 100% smoke-free legislation may be more effective than partial bans, where there can be designated smoking areas or other exemptions, thus leaving social cues in place and missing an opportunity to hinder the social desirability of smoking. Similar to the smoking ban in bars, people who were allowed to smoke anywhere in their home had lower odds of being a Tryhard versus those who were not allowed to smoke anywhere in their home. Perhaps this suggests a role for legislation banning smoking in public housing as well as a role for partners and other people in the home to enforce a smoke-free household to the extent possible.

We also called attention to policy implications having the potential to assist current everyday smokers to successfully quit smoking. Nagelhout et. al addressed smoking bans in European countries. In this article, they found that in England, after a comprehensive smoking ban, English 47.3% smokers were successful in their quit attempts, compared to 26.4% before the legislation was enacted. In the Netherlands however, where there was not a comprehensive ban, there was no statistically significant difference in the rate of successful attempts to quit smoking before and after the legislation, perhaps suggesting that a full ban is necessary to have the smoking not be socially desirable [51]. Given experience from other countries, it seems plausible that a full smoking ban in public places in the US will help

people successfully quit smoking.

5.7 Conclusion

This manuscript shows that useful insights can follow from using Fuzzy Forests to identify important predictors in a predictive model setting coupled with weighted logistic regression once predictors have been identified. The analysis also found that there is merit in using survey weights to adjust for bias in estimates and standard errors from traditional methods such as logistic regression. The comparison between Fuzzy Forests with and without inclusion of the self-response weights indicates how important it also is to adjust for them in machine learning methods as well. Thus it is not advisable to ignore the survey design and naively apply machine learning methods to data, even if they are just doing variable screening.

Table 5.1: Demographic characteristics

Variable		Overall (N =876)
		Weighted Mean(CI)
Age		40.43(38.94, 41.93)
		N(%)
Marital Status		
	Married	334(35.77)
	Widowed	28(3.10)
	Divorced	214(23.93)
	Separated	48(6.06)
	Never Married	252(31.15)
Gender		
	Female	769(86.64)
	Male	107(13.36)
Military(Ever Active Duty Armed Forces)		
	Yes	31(3.73)
	No	845(96.27)
Education		
	HS/GED or Lower	305(35.88)
	Some college(no degree)	203(23.98)
	Associate Degree	255(27.32)
	Bachelor Degree	80(8.92)
	Masters Degree	18(2.06)
	Professional Degree or doctorate	15(1.84)
Race		
	White	716(80.19)

Table 5.1 continued from previous page

	Black	106(14.11)
	Other	54(5.7)
Hispanic		
	Yes	44(6.75)
	No	832(93.25)
Country of Birth		
	US	835(95.39)
	Foreign Country	41(4.61)
Profession		
	Clinical laboratory technologists and technicians	20(1.88)
	Dental assistants	22(2.59)
	Dental hygienists	5(0.56)
	Diagnostic related technologists and technicians	30(3.41)
	Dietitians and nutritionists	3(0.34)
	Emergency medical technicians and paramedics	15(2.22)
	Health diagnosing and treating practitioner support technicians	49(5.6)
	Licensed practical and licensed vocational nurses	84(9.08)
	Medical assistants	14(1.76)
	Medical assistants and other healthcare support occupations	66(7.79)
	Medical records and health information technicians	21(2.29)
	Medical transcriptionists	1(0.1)
	Miscellaneous health technologists and technicians	14(1.49)
	Miscellaneous healthcare support occupations	9(0.99)
	Nurse anesthetists	2(0.27)
	Nursing, psychiatric, and home health aides	323(36.9)

Table 5.1 continued from previous page

Occupational therapists	6(0.85)
Opticians, dispensing	5(0.76)
Optometrists	1(0.01)
Other healthcare practitioners and technical occupations	4(0.34)
Pharmacists	5(0.56)
Pharmacy aides	1(0.15)
Phlebotomists	1(0.23)
Physical therapist assistants and aides	6(0.84)
Physical therapists	6(0.5)
Physician assistants	1(0.04)
Physicians and surgeons	4(0.5)
Recreational therapists	1(0.22)
Registered nurses	138(15.18)
Respiratory therapists	9(1.27)
Speech-language pathologists	1(0.07)
Therapists, all other	9(1.21)

Table 5.2: Unweighted Fuzzy Forests by quitting resolution

	Diehard	Tryhard
Common Variables	how many hours worked, or reason for not working how many years did you smoke everyday(age - age started smoking everyday) age occupation code for primary job FULL/PART-TIME WORK STATUS when you smoked the most, how many cigarettes did you smoke each day industry code for primary job how many hours per week do you usually work at your main job	
Unique Variables	Medical doctor or dentist tells you to quit smoking in the past 12 months	In bars, cocktail lounges, and clubs, do you THINK that smoking SHOULD be allowed in all areas, some areas or not at all
	How long have you smoked everyday	On average, how many cigarettes do you now smoke each day
	Principal city/balance status	Usually buy cigarettes by the pack or carton
	Father's country of birth	Usual hours worked weekly

Table 5.2 continued from previous page

	Around this time 12 months ago, on average how many cigs did you smoke each day	Last week how many hrs did you actually work at your job
	Highest level of school completed or degree received	Around this time 12 months ago, on average how many cigs did you smoke each day
	Metropolitan area size	12 months ago, were you usually smoking menthol or non-menthol cigarettes
Time	275.31	335.65
OOB error rate	31.51%	34.82%

Table 5.3: Weighted Fuzzy Forests by quitting resolution

	Diehard	Tryhard
Common Variables	How soon after waking do you typically smoke your first cigarette of the day(min)	
	Metropolitan area size	
	Age	
	Highest level of school completed or degree received	
	Industry code for primary job	
	Occupation code for primary job	
	Presence of own children<18 yrs old by age group	
	How many hours worked, or reason for not working	
	How many years did you smoke everyday (age - age started smoking everyday)	
	In the past 12 months, a medical doctor or dentist told you to quit smoking	

Table 5.3 continued from previous page

	Smoking rules inside home	
	Age that you started smoking cigarettes everyday	
	Full / part-time work status	
Unique Variables	Family income	Usual hours worked weekly
	Marital status based on armed forces participation	Mainly work indoors or outdoors
	Mainly work indoors or outdoors	Usually buy cigarettes by the pack or carton
	Living quarters situation	In bars, cocktail lounges, and clubs, do you THINK that smoking SHOULD be allowed in all areas, some areas or not at all
	How long have you smoked everyday	On average, about how many cigs do you smoke each day
	When you smoked the most, how many cigarettes did you smoke each day	Age when you first started smoking cigarettes fairly regularly
	Principal city/ balance status	When you smoked the most, how many cigarettes did you smoke daily
Run time	1643.2	1532.87
Average OOB error	0.089	0.112
Average AUC	0.957	0.953

Table 5.4: Weighted Logistic for Diehards

Variable	Beta	Odds Ratio (CI)	Pvalue
When you smoked the most, how many cigs did you use daily	-0.04	0.96(0.94,0.99)	0.002
Smoke everyday			
Didn't smoke everyday (for all the years that you smoked)	-0.81	0.45(0.28,0.7)	0.0005
Smoked everyday for nearly (for all the years that you smoked)	−−	−−	−−
Medical doctor / dentist told you to quit in the past 12 months			
Didn't see doctor in past year	0.76	2.14(0.85,5.39)	0.1
Yes	-0.04	0.96(0.37,2.46)	0.93
No	−−	−−	−−
Smoking allowed in your home:			
Anywhere	0.62	1.86(1.21,2.85)	0.01
Some places	0.36	1.44(0.81,2.56)	0.22
Nowhere	−−	−−	−−
Race			
White	-0.47	0.62(0.26,1.53)	0.3
Black	-0.43	0.65(0.26,1.64)	0.36
Other	−−	−−	−−
Age first started smoking cigarettes fairly regularly	-0.01	0.99(0.95,1.03)	0.54
Gender			
Female	0.18	1.19(0.65,2.21)	0.57
Male	−−	−−	−−

Table 5.5: Weighted Logistic for Tryhards

Variable	Beta	Odds Ratio(CI)	Pvalue
In bars/cocktail lounges/clubs smoking is allowed in			
All areas	-0.6	0.55(0.31,0.99)	0.05
Some areas	-0.49	0.61(0.39,0.97)	0.04
No areas	−−	−−	−−
Smoking allowed in your home			
Anywhere	-0.48	0.62(0.41,0.93)	0.02
Some places	-0.39	0.68(0.43,1.08)	0.1
Nowhere	−−	−−	−−
Race			
White	0.23	1.25(0.55,2.84)	0.59
Black	0.09	1.1(0.4,3.04)	0.86
Other	−−	−−	−−
Age first started smoking cigarettes fairly regularly	0.002	1(0.97,1.04)	0.93
Gender			
Female	-0.37	0.69(0.4,1.18)	0.17
Male	−−	−−	−−

Figure 5.1: Average Variable Importance from Weighted Fuzzy Forests for Diehards

Figure 5.2: Variable Importance from Weighted Fuzzy Forests for Tryhards

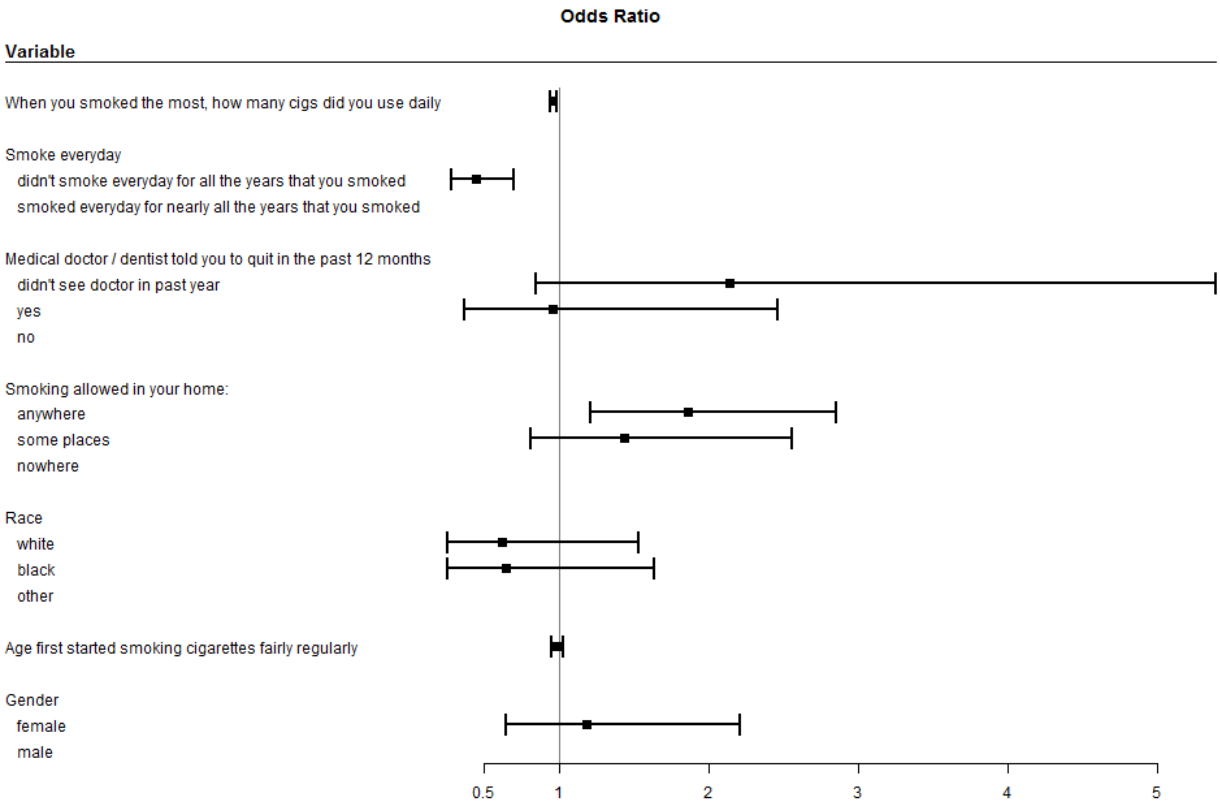

Figure 5.3: Weighted logistic odds ratio for Diehards

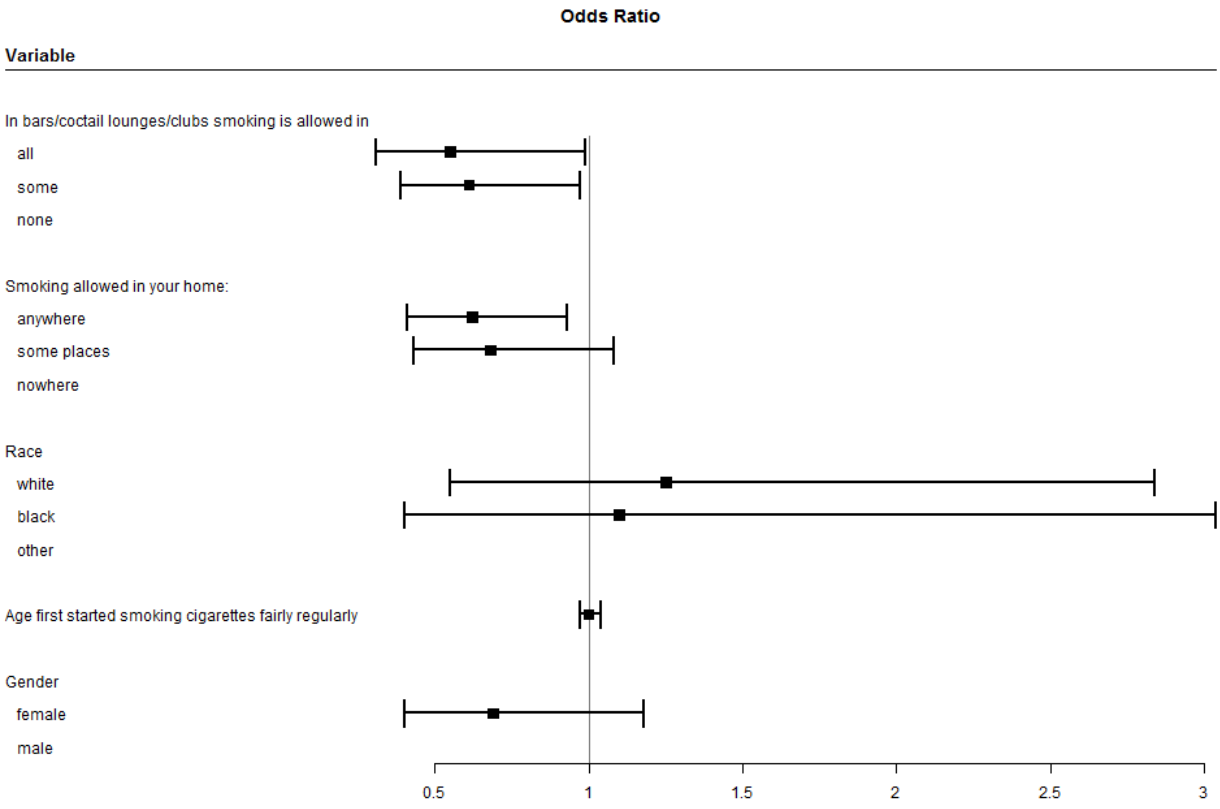

Figure 5.4: Weighted logistic odds ratio for Tryhards

CHAPTER 6

Conclusion

The exploration of homeless male ex-offenders, lithium responders to bipolar disorder, and factors associated with being a Diehard or Tryhard smoker all highlighted different facets of decision tree methods, with a particular focus on Fuzzy Forests. Since Fuzzy Forests are better suited to analyzing data that contains correlated predictors than Random Forests and CART, this method was used in all three analyses.

The results from these applications are mixed. In the bipolar analysis, Fuzzy Forests was not able to show any reproducible results between the retrospective and the prospective datasets. However, this is not that surprising given the limited sample size. It illustrates succinctly that machine learning is not the panacea to an under-powered study The homeless male ex-offender datasets show that among the decision tree methods; CART, Random Forests and Fuzzy Forests, the results are comparable, but with Fuzzy Forests being slightly more preferable in the testing sample. Lastly while the smoking data analysis does not compare decision tree methods, the question of whether to address the impact of utilizing the sampling weights to some degree or ignoring them is explored. The results indicate that while Fuzzy Forests, like all decision tree methods, is able to explore the relationships between all the predictors and the outcome, the overall accuracy is impacted when these weights are not incorporated in the analysis in some way. Overall through these various applications, the utility of Fuzzy Forests has been shown and the various facets of applying this method have been illustrated.

BIBLIOGRAPHY

[1] Smoke-free public housing and multifamily properties. https://www.hud.gov/program_offices/healthy_homes/smokefree. Accessed: 2010-04-17.

[2] Hierarchical clustering. Website. http://www.saedsayad.com/clustering_hierarchical.htm.

[3] Smokefree air laws. https://www.lung.org/our-initiatives/tobacco/smokefree-environments/smokefree-air-laws.html.

[4] What is tus-cps? https://cancercontrol.cancer.gov/brp/tcrb/tus-cps/, .

[5] Current population survey, january 2011. tobacco use file. http://www.nber.org/cps/cpsjan11.pdf, .

[6] *Random Forests: Leo Breiman and Adele Cutler*, accessed June 20, 2017. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

[7] *Bipolar Disorder*, accessed June 20, 2019. https://www.nami.org/learn-more/mental-health-conditions/bipolar-disorder.

[8] *Deoxyribonucleic Acid (DNA) Fact Sheet*, accessed June 27, 2019. https://www.genome.gov/about-genomics/fact-sheets/Deoxyribonucleic-Acid-Fact-Sheet.

[9] *Bipolar Disorder*, accessed May 22, 2017. http://www.nimh.nih.gov/health/topics/bipolar-disorder/index.shtml.

[10] *Bipolar Disorder among adults*, accessed May 22, 2017. https://www.nimh.nih.gov/health/statistics/prevalence/bipolar-disorder-among-adults.shtml.

[11] *Genetic Association Information Network (GAIN) (n.d.)*, accessed May 25, 2017. https://www.genome.gov/19518664/.

[12] *California Department of Corrections and Rehabilitation; Office of Research. 2015 outcome evaluation report*, August 2016. `http://www.cdcr.ca.gov/Adult_Research_Branch/Research_Documents/2015_Outcome_Evaluation_Report_8-25-2016.pdf`.

[13] Charu C Aggarwal. *Data classification: algorithms and applications*. CRC press, 2014.

[14] Christophe Ambroise and Geoffrey J McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562–6566, 2002.

[15] Kellie J Archer and Ryan V Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, 2008.

[16] Stephen Babb. Quitting smoking among adults —united states, 2000 —2015. *MMWR. Morbidity and Mortality Weekly Report*, 65, 2017.

[17] Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Scientific American*, 288(5):50–59, 2003.

[18] Charles Barrios, Paul J Goodnick, and Tanveer A Chaudhry. Rapid cycling bipolar disorder. *Expert Opinion on Pharmacotherapy*, 2(12):1963–1973, 2001.

[19] Sven Bergmann, Jan Ihmels, and Naama Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biology*, 2(1):E9–E9, 2004.

[20] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[21] Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[22] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[23] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[24] Joanne E Callinan, Anna Clarke, Kirsten Doherty, and Cecily Kelleher. Legislative smoking bans for reducing secondhand smoke exposure, smoking prevalence and tobacco consumption. *Cochrane Database of Systematic Reviews*, (4), 2010.

[25] Adem Can, Thomas G Schulze, and Todd D Gould. Molecular actions and clinical pharmacogenetics of lithium therapy. *Pharmacology Biochemistry and Behavior*, 123: 3–16, 2014.

[26] Stephen J Chanock, Teri Manolio, Michael Boehnke, Eric Boerwinkle, David J Hunter, Gilles Thomas, Joel N Hirschhorn, Goncalo Abecasis, David Altshuler, Joan E Bailey-Wilson, et al. Replicating genotype —phenotype associations. *Nature*, 447(7145):655, 2007.

[27] Daniel Conn and Christina M. Ramirez. *Random Forests and Fuzzy Forests in Biomedical Research*, page 168196. Analytical Methods for Social Research. Cambridge University Press, 2016. doi: 10.1017/CBO9781316257340.008.

[28] Daniel Conn, Tuck Ngun, Gang Li, and Christina Ramirez. Fuzzy forests: Extending random forests algorithm for correlated, high-dimensional data. *Journal of Statistical Software*.

[29] Nick Craddock and Ian Jones. Genetics of bipolar disorder. *Journal of Medical Genetics*, 36(8):585–594, 1999.

[30] David Curtis, Gursharan Kalsi, Jon Brynjolfsson, Melvin McInnis, Jane O'Neill, Ciaran Smyth, Eamonn Moloney, Patrice Murphy, Andrew McQuillin, Hannes Petursson, et al. Genome scan of pedigrees multiply affected with bipolar disorder provides further support for the presence of a susceptibility locus on chromosome 12q23-q24, and suggests the presence of additional loci on 1p and 1q. *Psychiatric Genetics*, 13(2):77–84, 2003.

[31] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3, 2006.

[32] Thomas G Dietterich. Machine-learning research: Four current directions. *AI Magazine*, 18(4):97–136, 1997.

[33] Bradley Efron and Robert Tibshirani. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

[34] Julie Garnham, Alana Munro, Claire Slaney, Marsha MacDougall, Michael Passmore, Anne Duffy, Claire O'Donovan, Andrew Teehan, and Martin Alda. Prophylactic treatment response in bipolar disorder: results of a naturalistic observation study. *Journal of Affective Disorders*, 104(1-3):185–190, 2007.

[35] Elaine Green, Gareth Elvidge, Nick Jacobsen, Beate Glaser, Ian Jones, Michael C ODonovan, George Kirov, Michael J Owen, and Nick Craddock. Localization of bipolar susceptibility locus by molecular genetic analysis of the chromosome 12q23 –q24 region in two pedigrees with bipolar disorder and dariers disease. *American Journal of Psychiatry*, 162(1):35–42, 2005.

[36] Greg A Greenberg and Robert A Rosenheck. Homelessness in the state and federal prison population. *Criminal Behaviour and Mental Health*, 18(2):88–103, 2008.

[37] Tiffany A Greenwood, Hagop S Akiskal, Kareen K Akiskal, Bipolar Genome Study, and John R Kelsoe. Genome-wide association study of temperament in bipolar disorder reveals significant associations with three novel loci. *Biological Psychiatry*, 72(4):303–310, 2012.

[38] Baptiste Gregorutti, Bertrand Michel, and Philippe Saint-Pierre. Correlation and variable importance in random forests. *arXiv preprint arXiv:1310.5726*, 2013.

[39] Paul Grof, Martin Alda, Eva Gror, Petr Zvolsky, and Mary Walsh. Lithium response and genetics of affective disorders. *Journal of Affective Disorders*, 32(2):85–95, 1994.

[40] Donna L Hall, Richard P Miraglia, Li-Wen G Lee, Deborah Chard-Wierschem, and Donald Sawyer. Predictors of general and violent recidivism among smi prisoners returning

to communities in new york state. *Journal of the American Academy of Psychiatry and the Law Online*, 40(2):221–231, 2012.

[41] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[42] Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. New York, N.Y. Springer, 2009. ISBN 978-0-387-84857-0.

[43] Steve Horvath. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. Springer Science & Business Media, 2011.

[44] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15 (3), 2006.

[45] Ahmed Jamal. Current cigarette smoking among adults —united states, 2005 —2015. *MMWR. Morbidity and Mortality Weekly Report*, 65, 2016.

[46] N Kleindienst, RR Engel, and W Greil. Which clinical factors predict response to prophylactic lithium? a systematic review for bipolar disorders. *Bipolar Disorders*, 7 (5):404–417, 2005.

[47] Andy Liaw and Matthew Wiener. Package randomforest: Breiman and cutlers random forests for classification and regression. *CRAN Reference manual*, 2015.

[48] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[49] João Maroco, Dina Silva, Ana Rodrigues, Manuela Guerreiro, Isabel Santana, and Alexandre de Mendonça. Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, 4(1):299, 2011.

[50] Michael J McCarthy, Susan G Leckband, and John R Kelsoe. Pharmacogenetics of lithium response in bipolar disorder. *Pharmacogenomics*, 11(10):1439–1465, 2010.

[51] Gera E Nagelhout, Hein de Vries, Christian Boudreau, Shane Allwright, Ann McNeill, Bas van den Putte, Geoffrey T Fong, and Marc C Willemsen. Comparative impact of smoke-free legislation on smoking cessation in three european countries. *The European Journal of Public Health*, 22(suppl_1):4–9, 2012.

[52] Kristin Nicodemus, James Malley, Carolin Strobl, and Andreas Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11(1):110, 2010.

[53] Kristin K. Nicodemus and James D. Malley. Predictor correlation impacts machine learning algorithms: Implications for genomic studies. *Bioinformatics*, 25(15):1884–1890, 2009.

[54] Kristin K Nicodemus and James D Malley. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, 25(15):1884–1890, 2009.

[55] Stephanie Nissen, Sherri Liang, Tatyana Shehktman, John R Kelsoe, and Bipolar Genome Study (BiGS). Evidence for association of bipolar disorder to haplotypes in the 22q12. 3 region near the genes stargazin, ift27 and parvalbumin. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 159(8):941–950, 2012.

[56] Adeline Nyamathi, Benissa E Salem, David Farabee, Elizabeth Hall, Sheldon Zhang, Mark Faucette, Doug Bond, and Kartik Yadav. Impact of an intervention for recently released homeless offenders on self-reported re-arrest at 6 and 12 months. *Journal of Addictive Diseases*, 36(1):60–71, 2017.

[57] US Department of Health and Human Services. The health consequences of smoking —50 years of progress: A report of the surgeon general, 2014.

[58] Joan Petersilia. *When prisoners come home: Parole and prisoner reentry*. Oxford University Press, 2003.

[59] Joan Petersilia. Hard time: Ex-offenders returning home after prison. *Corrections Today*, 67(2):66–71, 2005.

[60] MÁRTON PÓSFAI, GABRIELE MUSELLA, MAURO MARTINO, NICOLE SAMAY, SARAH MORRISON, AMAL HUSSEINI, and PHILIPP HOEVEL. Network science.

[61] Suggests RColorBrewer and Maintainer Andy Liaw. Package randomforest. 2018.

[62] Stephan Ripke, Benjamin M Neale, Aiden Corvin, James TR Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, David A Collier, Hailiang Huang, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511 (7510):421–427, 2014.

[63] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[64] Linda Sarna, Stella Aguinaga Bialous, Karabi Sinha, Qing Yang, and Mary Ellen Wewers. Are health care providers still smoking? data from the 2003 and 2006/2007 tobacco use supplement-current population surveys. *Nicotine & Tobacco Research*, 12(11):1167–1171, 2010.

[65] Linda Sarna, Stella Aguinaga Bialous, Karabi Nandy, Anna Liza Malazarte Antonio, and Qing Yang. Changes in smoking prevalences among health care professionals from 2003 to 2010-2011. *JAMA*, 311(2):197–199, 2014.

[66] Anthony Scime and Gregg R Murray. Social science data analysis: The ethical imperative. In *Ethical data mining applications for socio-economic development*, pages 131–147. IGI Global, 2013.

[67] Mark Robert Segal. Regression trees for censored data. *Biometrics*, pages 35–47, 1988.

[68] Nong Shang and Leo Breiman. Distribution based trees are more accurate. *Ionosphere*, 2(33):351, 1996.

[69] Erin N Smith, Daniel L Koller, Corrie Panganiban, Szabolcs Szelinger, Peng Zhang, Judith A Badner, Thomas B Barrett, Wade H Berrettini, Cinnamon S Bloss, William Byerley, et al. Genome-wide association of bipolar disorder suggests an enrichment of replicable associations in regions near genes. *PLoS Genetics*, 7(6):e1002134, 2011.

[70] Alexander Statnikov, Lily Wang, and Constantin F Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9(1):319, 2008.

[71] Dan Steinberg and Phillip Colla. Cart: classification and regression trees. *The top ten algorithms in data mining*, 9:179, 2009.

[72] Carolin Strobl. *Statistical Issues in Machine Learning - Towards Reliable Split Selection and Variable Importance Measures*. PhD thesis, Institut fur Statistik, Ludwig Maximilian University Munich, Munich, Germany, May 2008.

[73] Carolin Strobl, Anne-Laure Boulesteix, and Thomas Augustin. Unbiased split selection for classification trees based on the gini index. *Computational Statistics & Data Analysis*, 52(1):483–501, 2007.

[74] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1): 307, 2008.

[75] Carolyn Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9: 307, 2008.

[76] Girija Syamlal, Jacek M Mazurek, Eileen Storey, and Shanta R Dube. Cigarette smoking prevalence among adults working in the health care and social assistance sector,

2008 to 2012. *Journal of occupational and environmental medicine/American College of Occupational and Environmental Medicine*, 57(10):1107, 2015.

[77] Paul M Thompson, Jason L Stein, Sarah E Medland, Derrek P Hibar, Alejandro Arias Vasquez, Miguel E Renteria, Roberto Toro, Neda Jahanshad, Gunter Schumann, Barbara Franke, et al. The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging and Behavior*, 8(2):153–182, 2014.

[78] Brian TD Tobe, Andrew M Crain, Alicia M Winquist, Barbara Calabrese, Hiroko Makihara, Wen-ning Zhao, Jasmin Lalonde, Haruko Nakamura, Glenn Konopaske, Michelle Sidor, et al. Probing the lithium-response pathway in hipscs implicates the phosphoregulatory set-point for a cytoskeletal modulator in bipolar pathogenesis. *Proceedings of the National Academy of Sciences*, 114(22):E4462–E4471, 2017.

[79] L Tondo, J Hennen, and RJ Baldessarini. Rapid-cycling bipolar disorder: effects of long-term treatments. *Acta Psychiatrica Scandinavica*, 108(1):4–14, 2003.

[80] Jeremy Travis. *But they all come back: Facing the challenges of prisoner reentry*. The Urban Insitute, 2005.

[81] Eduard Vieta and Jose Sanchez-Moreno. Acute and long-term treatment of mania. *Dialogues in Clinical Neuroscience*, 10(2):165, 2008.

[82] Teresa W Wang, Kat Asman, Andrea S Gentzke, Karen A Cullen, Enver Holder-Hayes, Carolyn Reyes-Guzman, Ahmed Jamal, Linda Neff, and Brian A King. Tobacco product use among adults –united states, 2017. *Morbidity and Mortality Weekly Report*, 67(44): 1225–1232, 2018.

[83] Naomi R Wray, Stephan Ripke, Manuel Mattheisen, Maciej Trzaskowski, Enda M Byrne, Abdel Abdellaoui, Mark J Adams, Esben Agerbo, Tracy M Air, Till MF Andlauer, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, 50(5):668–681, 2018.

[84] Momiao Xiong, Wuju Li, Jinying Zhao, Li Jin, and Eric Boerwinkle. Feature (gene) selection in gene expression-based tumor classification. *Molecular Genetics and Metabolism*, 73(3):239–247, 2001.

[85] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[86] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[87] Heping Zhang and Burton H Singer. *Recursive partitioning and applications*. Springer, 2010.